UNIVERSIDADE FEDERAL DO PARANÁ

KÁTIA DE PAIVA LOPES

IDENTIFICAÇÃO DE *SPLICING* ALTERNATIVO EM SÍTIOS DE INÍCIO DE TRANSCRIÇÃO DE mRNAS DE ARROZ UTILIZANDO DADOS DE RNA-Seq

CURITIBA

2012

KATIA DE PAIVA LOPES

IDENTIFICAÇÃO DE *SPLICING* ALTERNATIVO EM SÍTIOS DE INÍCIO DE TRANSCRIÇÃO DE mRNAS DE ARROZ UTILIZANDO DADOS DE RNA-Seq

Dissertação apresentada ao Programa de Pós-Graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, da Universidade Federal do Paraná, como requisito parcial para a obtenção do título de Mestre em Bioinformática.

Orientador: Prof. Dr. Adriano Barbosa da Silva.

Co-orientador: Prof. Dr. Emanuel Maltempi de Souza.

CURITIBA

2012

LOPES, Kátia de Paiva

Identificação de *Splicing* alternativo em sítios de início de transcrição de mRNAS de arroz atualizando dados de RNA-Seq / Kátia de Paiva Lopes; orientador, Adriano Barbosa da Silva; coorientador, Emanuel Maltempi de Souza. - Curitiba, PR, 2012.

125 f.

Dissertação (Mestrado) - Universidade Federal do Paraná, Setor de Educação Profissional e Tecnológica. Programa de Pós-Graduação em Bioinformática.

Inclui referências

1. RNA-Seq. 2. Bioinformática. 3. Ciência da Computação. I. Silva, Adriano Barbosa da. II. Souza, Emanuel Maltempi de. III. Universidade Federal do Paraná, Setor de Educação Profissional e Tecnológica. IV. Título.

TERMO DE APROVAÇÃO

KATIA DE PAIVA LOPES

Identificação de *splicing* alternativo e Sítios de Início de Transcrição de mRNAs de arroz utilizando dados de RNA-Seq

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre em Bioinformática, pelo Programa de Pós-graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, da Universidade Federal do Paraná, pela seguinte banca examinadora:

Orientador: Prof Dr Adriano Barbosa da Silva

Coorientador: Prof Dr Emanuel Maltempi de Souza

Dr. Christian Macagnan Probst
Fundação Oswaldo Cruz - Fio Cruz/PR

Drª. Michelle Zibetti Tadra Sfeir
Universidade Federal do Paraná - UFPR

Curitiba, 19 de dezembro de 2012

Dedico este trabalho aos meus pais.

AGRADECIMENTOS

Primeiramente a Deus pela saúde e aos meus pais pelo incentivo. Também aos meus orientadores, Prof. Dr. Emanuel Maltempi de Souza pelo direcionamento do trabalho e por compartilhar sua profunda experiência profissional. Ao Prof. Dr. Adriano Barbosa da Silva pela ajuda em todas as etapas do trabalho e pelo direcionamento da pesquisa tanto na parte biológica quanto computacional. Alguém que além de tudo foi um amigo.

Aos demais professores do Programa, ao Prof. Lucas Ferrari por ter cedido os servidores do seu projeto para tarefas relacionadas ao trabalho no início do curso. Aos Professores Doutores Fábio de Oliveira Pedrosa, Roseli Wassem e a Doutoranda Liziane Cristina pela disponibilização dos dados, troca de experiências, acompanhamento nos experimentos em laboratório e por compartilhar o tão precioso tempo.

A todos os colegas de curso das turmas de 2010, 2011 e 2012 sem exceção. Em especial ao Rafael Covre pela ajuda também em todas as etapas do trabalho. À Kelly Rafaela, William Joanico, Leviston da Silveira, Sérgio Fujji, Bárbara Nobre e Ricardo Vialle pela disposição em ajudar nas disciplinas e pela compreensão nos momentos estressantes. O mesmo eu digo para os colegas novos Fabiano Gomes, Vitor Cedran, Jeovane Alves e Bárbara Abreu.

Aos meus familiares, avós, tios e primos que apoiaram. Em especial aos meus primos Sueli e Roberto que me permitiram morar com eles por nove meses no Paraná e aos meus tios Maria e Geraldo por todo apoio. Às minhas primas Solange, Sônia e Alessandra que estiveram comigo durante todo o curso. Ao meu namorado Sérgio pelo apoio e por entender que nem sempre eu pude estar presente.

Aos meus amigos de Minas Gerais que se encontraram comigo todas as vezes que voltei para casa e deram o incentivo final. Entre eles Carla, Ederson, Jorge, Nivaldo, Tilibra, Rafael, Soneca, Faustim, Adriana, Marcos, Felipe, Paula e Gustavo Faria, minhas irmãs Kelen, Karina e à pequena Sofia. A Diego Araújo que me impulsionou a vir e ao meu Orientador de Iniciação Científica Prof. Dr. Sandro Renato Dias por todo apoio e ouvidos.

Ao Programa de Pós-Graduação em Bioinformática da UFPR pela oportunidade e pela utilização do cluster. Aos Professores Doutores Maria Berenice, Jeroniza Marchaukoski, Roberto Tadeu Raittz e ao Doutorando Dieval Guizelini por todo apoio prestado. Às Secretárias Suzana e Léa pela amizade e disposição em ajudar sempre.

E por fim, aos órgãos financiadores da bolsa de estudo e fomento para participação em eventos, sem os quais eu não poderia ter me dedicado integralmente a esse Mestrado: CNPq, CAPES, INCT-FBN, PRPPG, UFPR e SEPT. Ao CESUP (Centro Nacional de Super Computação) pela disponibilização de uma conta no cluster SGI-Altix, O Gauss.

Seio de Minas

Eu nasci no celeiro da arte
No berço mineiro
Sou do campo, da serra
Onde impera o minério de ferro
Eu carrego comigo no sangue
um dom verdadeiro
De cantar melodias de Minas
No Brasil inteiro

Sou das Minas de ouro
Das montanhas Gerais
Eu sou filha dos montes
E das estradas reais
Meu caminho primeiro
Vem brotar dessa fonte
Sou do seio de Minas
Nesse estado um diamante

Paula Fernandes

SUMÁRIO

LISTA DE FIGURAS

LISTA DE QUADROS E TABELAS

LISTA DE ABREVIATURAS E SIGLAS

cDNA	DNA sintetizado a partir de uma molécula de RNA mensageiro.
COG	*Clusters of Orthologous Groups of proteins.*
CR	Controle.
D.E	Diferencialmente expresso.
EST	*Expressed sequence tag*.
FBN	Fixação biológica de nitrogênio.
FPKM	*Fragments per kilobase of exon per million fragments mapped.*
GENOPAR	Consórcio criado para mapear o genoma do *Herbaspirillum seropedicae*, o GENOPAR foi estruturado como uma rede estadual de laboratórios composta por 12 unidades.
GO	*Gene Ontology.*
IGV	Software: *Integrative Genomics Viewer.*
IR	Inoculado.
IRGSP	*International Rice Genome Sequence Project* - Consórcio criado para análise do genoma de arroz.
NCBI	*National Center for Biotechnology Information*.
ORF	*Open read frame*. Fase aberta de leitura – Tradução livre.
pb	Pares de bases.
PCR	*Polymerase chain reaction*
RAP	*Rice Annotation Project* - Consórcio criado para anotação das sequências de arroz.
RAP-DB	Banco de dados com as sequências genômicas do arroz.
RNA-Seq	Sequenciamento de RNA.
RPKM	*Reads per kilobase per million mapped reads.*
TSS	*Transcription start site*. Sítio de início de transcrição.
UFPR	Universidade Federal do Paraná.
UFRGS	Universidade Federal do Rio Grande do Sul.

RESUMO

RNA-Seq é uma técnica que permite quantificar os níveis de expressão de uma forma muito mais precisa do que os métodos empregados anteriormente. Estudos que utilizaram esse método já alteraram a visão da extensão e da complexidade de transcriptomas eucarióticos. Portanto, para este trabalho, foram utilizadas plantas de arroz como organismo modelo, devido a sua importância global como cultura alimentar. Foram usadas sequências curtas de mRNA obtidas de plantas de arroz (*Oryza sativa* subs. *japonica*) inoculada ou não com a bactéria fixadora de nitrogênio *Herbaspirillum seropedicae* (Brusamarello-Santos *et al.*, dados não publicados). As sequências foram separadas de acordo com as amostras sequenciadas: dois grupos correspondem às amostras de três e sete dias após inoculação, composta por 328.409.635 de sequências de RNA-Seq trimadas. Assim, foi possível reportar o número de TSS's (*Transcriptional Start Sites*) e os TSS's diferencialmente expressos. Além disso, transcritos montados foram comparados com as estruturas gênicas dos genes anotados na base de dados RAP-DB (transcritos idênticos ou *splicing* alternativo) ou sem anotação (como os transcritos novos). Um total de 202.770.984 (61,7%) *reads* mapearam no genoma do arroz: 216.980.165 (75,2%) são *reads* exônicas, e 71.493.386 (24,8%) mapearam em junções de éxons. Esses números totalizam 288.473.551 de *reads* alinhadas, evidenciando que 85.702.567 dessas mapeavam em mais de um lugar. Um total de 6.942 (16,4%) e 4.915 (11,6%) genes obtiveram cobertura nas bibliotecas CR3 e IR3; e 6.733 (15,9%) e 3.807 (9%) genes para as bibliotecas CR7 e IR7, respectivamente. Posteriormente, foi possível reportar o número total de genes mapeados, anotados, não-anotados e transcritos com *splicing* alternativo para o genoma total, realizando a análise até mesmo por cromossomos em separado. Finalmente, após o alinhamento com o BLAST, 691 (3,31%) transcritos não-anotados obtiveram *match* com as proteínas do banco de dados NR. Em análises futuras, o restante, 20.185 (96,7%) transcritos não-anotados poderão ser comparados com as regiões de microRNA conhecidas no genoma do arroz. Aqueles que não mapearem nessas regiões devem ser sujeitos a uma análise de predição de estrutura secundária, a fim de verificar quanto a possíveis novas moléculas de microRNA reguladoras presentes neste conjunto de dados.

ABSTRACT

RNA-Seq allows the measurement of transcripts expression levels in a manner far more precise than previous methods. Studies using this approach have already altered our view of the extent and complexity of eukaryotic transcriptomes. However, in this work rice plants were used as model organism, due to its global importance as food crop. We used short mRNA sequences obtained for rice plants (*Oryza sativa* subsp. *japonica*) inoculated or not with the nitrogen-fixing bacteria *Herbaspirillum seropedicae* (Brusamarello-Campos *et al.,* unpublished data). Sequences have been separated according to the sequenced rice samples: two groups corresponding to samples three and seven days after inoculation, respectively, composed by 328,409,635 RNA-Seq trimmed *reads*. Hitherto, we report the amount of identified TSS's and TSS's differentially expressed. Finally, assembled transcripts have been checked by their (dis-)agreement with RAPDB gene structures (as for alternatively spliced transcripts) or lack of annotation (as for newly transcribed transcripts). A total of 202,770,984 (61.7%) *reads* have been mapped to rice genome: 216,980,165 (75.2%) as exonic *reads*, and 71,493,386 (24.8%) were aligned to exon junctions. These numbers sum up to 288,473,551 read alignments, evidencing that 85,702,567 alignment cases correspond to ambiguous mapping. A total of 6,942 (16.4%) and 4,915 (11.6%) genes were covered by *reads* in the datasets CR3d and IR3d; and, 6,733 (15.9%) and 3,807 (9%) genes for datasets CR7d and IR7d, respectively. Next, we report the total number of mapped, annotated, unannotated and alternatively spliced transcripts for the total genome, even the analysis have been performed for individual chromosomes. Finally, after BLAST alignment, 691 (3.31%) unannotated transcripts obtained matches to protein sequences from NR database. In future analysis, the remaining 20,185 (96.7%) unannotated transcripts will be compared to known microRNA spanning regions within the rice genome. Those not mapped to these regions will be subject to secondary structure prediction analysis in order to verify for possible new regulatory microRNA molecules represented in our dataset.

1. INTRODUÇÃO

1.1. *Oryza sativa* subs. *japonica*

Oryza sativa é o nome científico de uma das cultivares mais importantes do mundo: o arroz. Esta espécie pertence a divisão das Angiospermas (Magnoliophyta), subgrupo monocotiledôneas e à família *Poaceae*. O arroz é usado na nutrição de mais da metade da população global e desempenha um papel fundamental na nutrição humana por pelo menos 10.000 anos (INTERNATIONAL RICE GENOME PROJECT, 2005).

Dada sua importância comercial, em 1998 foi criado o consórcio *International Rice Genome Sequence Project* (IRGSP), que une recursos de 10 países para sequenciamento e estudo do genoma do arroz, subespécie *japonica* (INTERNATIONAL RICE GENOME PROJECT, 2005). Posteriormente, em 2004 foi criado o *Rice Annotation Project* (RAP) com o objetivo de permitir uma análise mais eficiente da informação genômica por meio do processo de anotação das sequências de arroz (RICE ANNOTATION PROJECT CONSORTIUM, 2008).

Os esforços e as análises destes consórcios resultaram em vários métodos e ferramentas que envolvem anotação automática de sequências de DNA, além de um banco de dados com o objetivo de coletar informações relevantes para bioinformática e genômica funcional, o RAP-DB, disponível em <http://rapdb.dna.affrc.go.jp/>. A seguir, são apresentadas algumas características marcantes do genoma do arroz, conforme disponibilizadas por estes consórcios:

- Como resultado da montagem do *draft* (inicial) das sequências de arroz apresentada pelo IRGSP, o tamanho do genoma foi estimado em 389 Mb. Um tamanho 260 Mb maior que a planta modelo *Arabidopsis thaliana*.
- Um total de 2.859 genes são únicos do arroz e outros cereais, importantes para diferenciar linhagens de monocotiledôneas e dicotiledôneas.
- O draft do IRGSP revela uma cobertura de 69% para a subespécie *indica* e 78% para *japonica* relativa às sequências baseadas em mapas cromossômicos.

- As sequências e anotação do genoma do arroz estão na versão 5, disponíveis no site do RAP-DB.
- Possui 43,6% de conteúdo GC.
- 12 cromossomos.
- 1 sequência de mitocôndria e 1 de cloroplasto representadas no banco de dados.
- 42.401 genes anotados e depositados no RAP-DB.
- 214.999 proteínas depositadas no NCBI[1].

(INTERNATIONAL RICE GENOME PROJECT, 2005)

(RICE ANNOTATION PROJECT CONSORTIUM, 2008)

Aproximadamente, 71% das proteínas preditas de arroz possuem proteínas homólogas em *Arabidopsis thaliana*. Reciprocamente, 89,8% das proteínas de *Arabidopsis* possuem homólogos no proteoma do arroz. Como esperado, um percentual menor é encontrado quando outras espécies são comparadas: 38,1% de proteínas homólogas em *Drosophila,* 40,8% em humanos, 36,5% em *Caenorhabditis elegans*, 30,2% em levedura, 17,6% em *Synechocystis* e 10,2% em *Escherichia coli* (INTERNATIONAL RICE GENOME PROJECT, 2005).

Na Figura 1 é possível visualizar o resultado da similaridade encontrada entre as proteínas de arroz e *Arabidopsis*. As proteínas foram classificadas em 5 grupos: homólogo do que não é planta (Non-plant), plantas sem flores (*Plant*), planta com flores (*Flowering plant*), monocotiledôneas e eudicotiledôneas (Monocot/Eudicot) e proteínas específicas de *Oryzeae* (tribo) e *Arabidopsis* (ITOH, TANAKA, *et al.*, 2007). O número de homólogos específicos de cada planta foi semelhante, apesar de haver vários transcritos encontrados especificamente para *Oryzeae* e *Arabidopsis.*

[1] NCBI: <http://www.ncbi.nlm.nih.gov/Taxonomy/> Acesso em 05 de setembro de 2012.

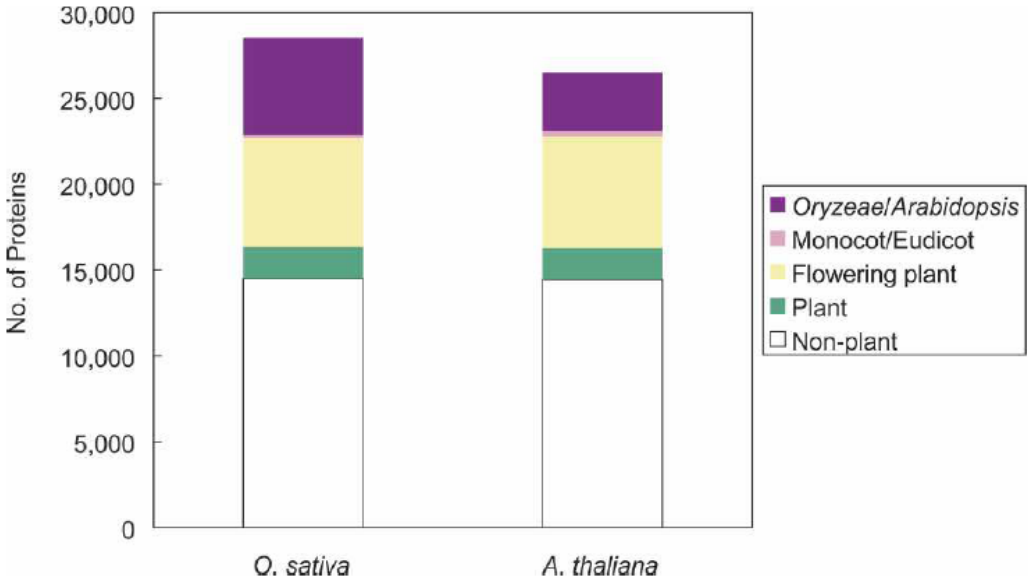


Figura 1 – Similaridade encontrada entre as proteínas de arroz e *Arabidopsis*.

As proteínas foram classificadas em 5 grupos: homólogo do que não é planta (Non-plant), plantas sem flores (Plant), planta com flores (Flowering plant), monocotiledôneas e eudicotiledôneas (Monocot/Eudicot) e proteínas específicas de Oryzeae (tribo) e Arabidopsis thaliana.

Fonte: (ITOH, TANAKA, *et al.*, 2007)

Como parte do estudo do genoma, foi possível gerar o mapa cromossômico do arroz (Figura 2). Na figura estão representados os 12 cromossomos, onde o mapa genético é mostrado no lado esquerdo e os *contigs* PAC/BAC[2] no lado direito. As distâncias genéticas são apresentadas em centimorgans (cM) e os mapas físicos são representados em tamanhos físicos relativos, ou seja, quantidade de pares de bases (INTERNATIONAL RICE GENOME PROJECT, 2005).

[2] PAC/BAC: Cromossomos artificiais bacterianos para clonagem dos fragmentos de arroz.

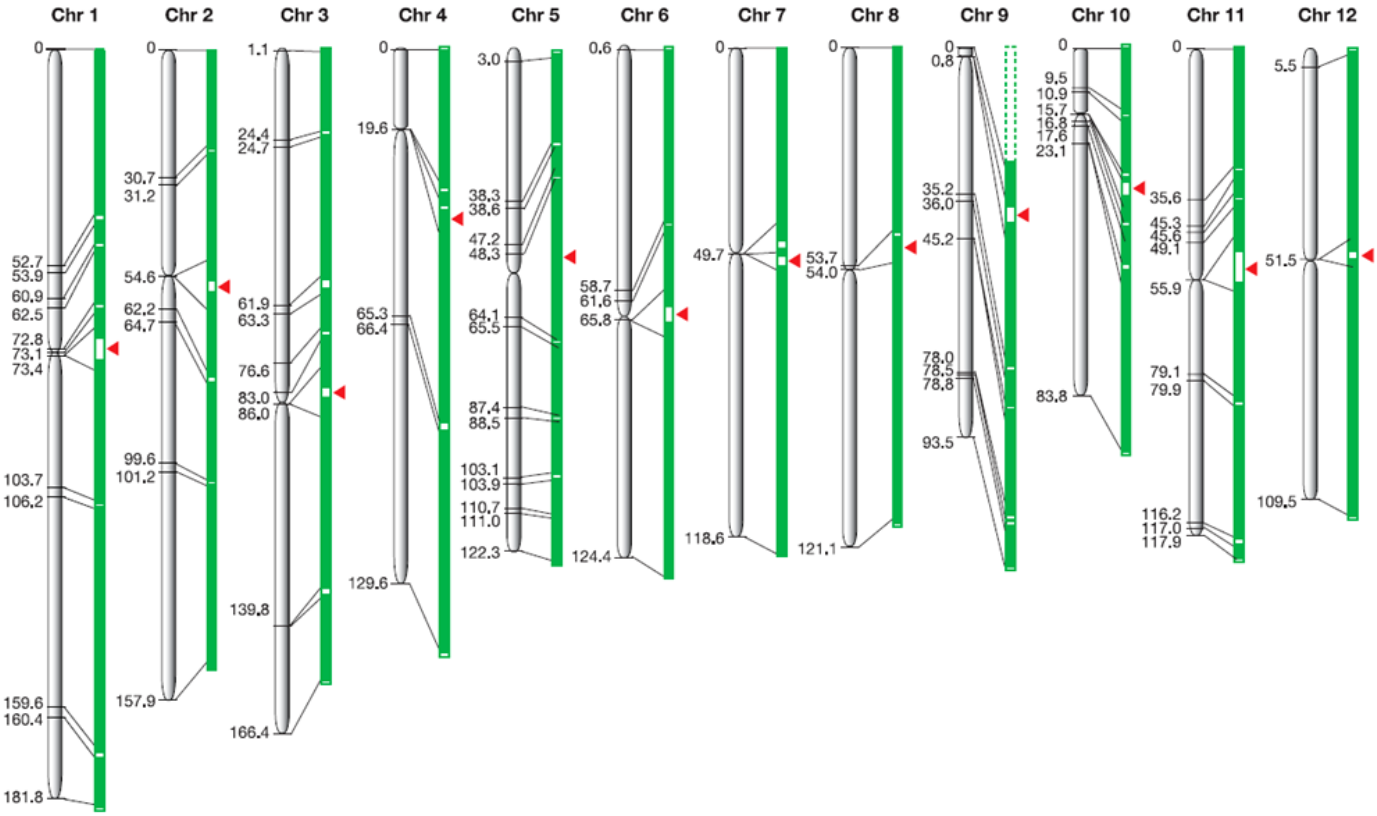


Figura 2- Mapa dos 12 cromossomos de arroz.

O mapa genético é mostrado no lado esquerdo e os contigs PAC/BAC no lado direito. A posição dos marcadores que flanqueiam o PAC/BAC são mostrados em verde. Lacunas físicas são mostradas em branco e a região organizadora do nucléolo, encontrada no cromossomo 9, é representada por uma linha tracejada. Constrições nos mapas genéticos e setas para a direita de mapas físicos representam as posições cromossômicas dos centrômeros. As distâncias genéticas são apresentadas em centimorgans (cM) e os mapas físicos são representados em tamanhos físicos relativos.

Fonte: (INTERNATIONAL RICE GENOME PROJECT, 2005).

O tamanho em pares de bases (pb) e Mega bases (Mb) de cada cromossomo baseado nos dados do sequenciamento estão apresentados na Tabela 1.

Tabela 1 - Tamanho dos cromossomos de arroz.

Cromossomo	Sequências (pb)	Total (Mb)
1	43.260.640	45,05
2	35.954.074	36,78
3	36.189.985	37,37
4	35.489.479	36,15
5	29.733.216	30,00
6	30.731.386	31,60
7	29.643.843	30,28

...continuação

Cromossomo	Sequências (pb)	Total (Mb)
8	28.434.680	28,57
9	22.692.709	30,53
10	22.683.701	23,96
11	28.357.783	30,76
12	27.561.960	27,77
Mitocôndria	497.529	0,50
Cloroplasto	136.447	0,14
Todos	371.367.432	389,46

Adaptação de: (INTERNATIONAL RICE GENOME PROJECT, 2005)

O estudo do genoma é uma abordagem eficaz para que seja encontrado um mecanismo capaz de melhorar a produção do arroz. Em 2005, estimava-se que a sua produção mundial deveria aumentar 30% nos 20 anos seguintes para atender a demanda populacional. Em contrapartida, a degradação ambiental, poluição e aumento de temperatura apresentam-se como restrições adicionais para aumento da produção (INTERNATIONAL RICE GENOME PROJECT, 2005). Portanto, torna-se vital aumentar o potencial de rendimento e estabilidade das plantas de arroz e uma maneira encontrada é a utilização de bactérias fixadoras de nitrogênio não patogênicas, conforme descrito nos itens a seguir.

1.2. *Herbaspirillum seropedicae*

Herbaspirillum seropedicae foi descrita por Baldani e colaboradores na década de 80, onde o nome *Herbaspirillum* se deve ao seu *habitat* ser localizado nas raízes de cereais, e *seropedicae*, refere-se ao local onde esta foi isolada pela primeira vez, em Seropédica, Rio de Janeiro, Brasil. Esta bactéria pode apresentar um, dois ou três flagelos, característica que lhe confere certa motilidade em seu micro-habitat. A Figura 3 apresenta uma micrografia eletrônica de *H. seropedicae,* onde três flagelos podem ser visualizados (BALDANI, SELDIN, *et al.*, 1986).

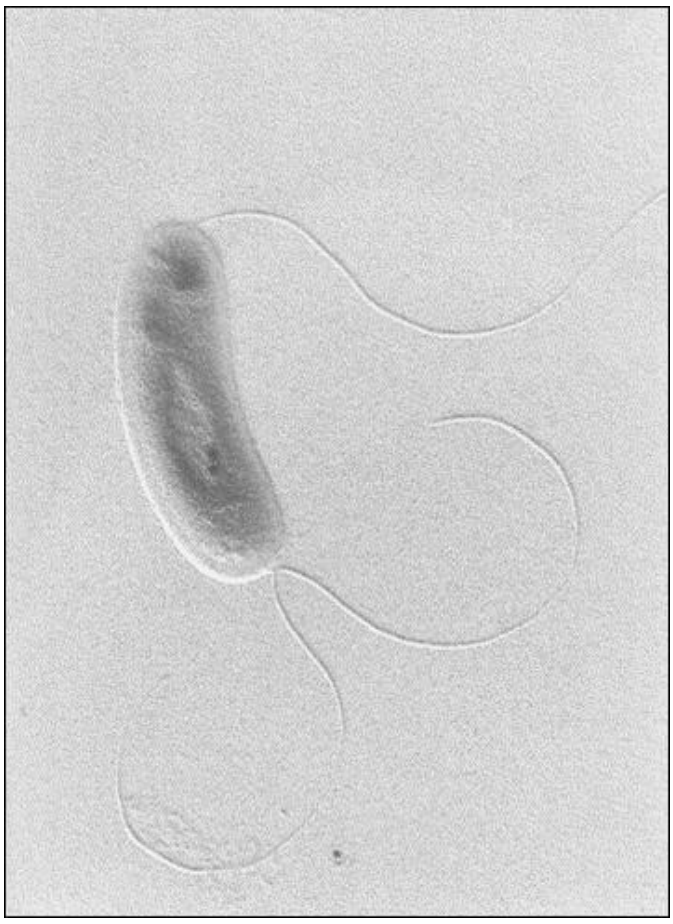


Figura 3 - Micrografia eletrônica da célula isolada de *H. seropedicae*.

A bactéria pode apresentar de um a três flagelos. Aqui representada apenas a bactéria com três flagelos.

Fonte: (BALDANI, SELDIN, *et al.*, 1986)

H. seropedicae é uma bactéria gram-negativa, endofítica, diazotrófica, capaz de colonizar gramíneas como arroz e cana-de-açúcar, aumentando sua produtividade em até 50%. A estirpe SmR1 foi anotada e sequenciada pelo Consórcio GENOPAR no Estado do Paraná. Seu genoma consiste em um cromossomo circular de 5.513.887 pb com 63,4% de conteúdo G+C. Possui 4.804 genes e um total de 4.735 possíveis ORF's (*Open Read Frame*). A sequência do genoma revelou que *H. seropedicae* é um organismo altamente versátil, com a capacidade de metabolizar uma grande variedade de fontes de carbono (PEDROSA, MONTEIRO, *et al.*, 2011). A Figura 4 mostra uma representação gráfica do seu genoma.

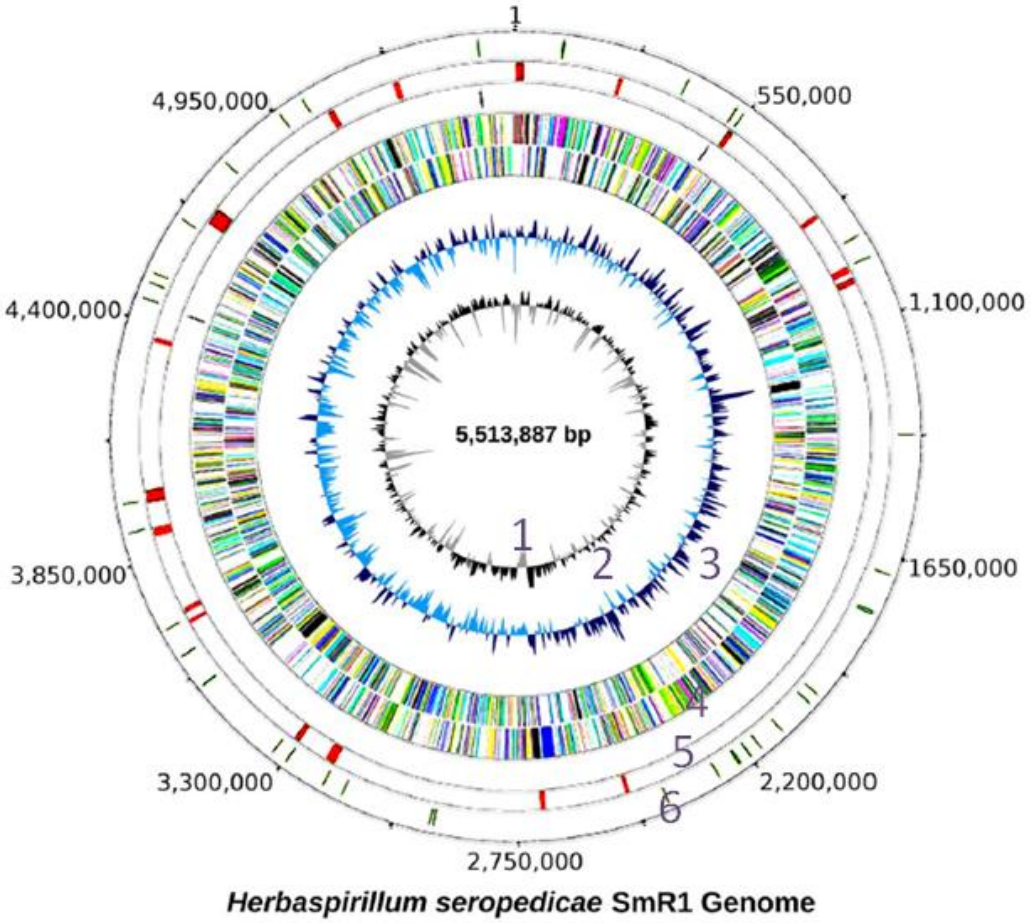


Figura 4 - Genoma do H. seropedicae.

1) Conteúdo G+C; 2) Correlação de G, GC: GC Skew; 3) Genes de acordo com suas categorias funcionais; 4) Operons de rRNAs; 5) Regiões transferidas horizontalmente; 6) Região idêntica à *Ricinus communis* (mamona).

Adaptada de: (PEDROSA, MONTEIRO, *et al.*, 2011).

Os genes fixadores de nitrogênio do *H. seropedicae* incluem *nifA, nifB, nifZ, nifZ1, nifH, nifD, nifK, nifE, nifN, nifX, nifQ, nifW, nifV, nifU* e *nifS* são encontrados em uma região genômica que abrange 37.547pb do genoma (PEDROSA, MONTEIRO, *et al.*, 2011). Estes genes são responsáveis pela fixação biológica de nitrogênio, descrita no item 1.2.1.

1.2.1 Fixação Biológica de Nitrogênio

Apesar de o Nitrogênio ser o elemento mais abundante do ar (~80% do ar atmosférico), os animais e as plantas não são capazes de metabolizá-lo em forma gasosa. Sendo assim, a Fixação Biológica de Nitrogênio (FBN) é realizada por bactérias diazotróficas e algumas cianobactérias. Já se sabe que, além das bactérias fixadoras de nitrogênio nas raízes das leguminosas, existem também as bactérias fixadoras de N2 endofíticas que atuam no interior de algumas plantas, como cana-de-açúcar, cereais e gramíneas forrageiras. Apesar de o produtor rural utilizar a adubação mineral para o fornecimento de nitrogênio às culturas agrícolas, esse adubo costuma ser caro e seu uso inadequado pode produzir impactos ambientais negativos (EMBRAPA, 2012). Dessa maneira a introdução das bactérias diazotróficas pode suprir a adubação mineral, onde se destaca as seguintes vantagens:

1. O menor uso de adubos nitrogenados, que resulta em economia para o produtor;
2. Contribuir para o auto-fornecimento do nitrogênio utilizado para a formação da planta minimiza os impactos do nitrogênio sobre o meio ambiente;
3. O uso de leguminosas como adubos verdes eficientes para FBN fornece nitrogênio para o solo e melhora suas propriedades físicas, químicas e biológicas;
4. A menor utilização da adubação química de nitrogênio tende a proporcionar uma melhoria da qualidade do solo, dos mananciais e de outros componentes ambientais.

(EMBRAPA, 2012)

Em linhas gerais, é possível dizer que com a FBN, há um aumento de produtividade da planta sem que haja maior área plantada, que proporciona inúmeros benefícios econômicos e ambientais. Para tanto, os sequenciadores de DNA de nova geração tem sido utilizados para decodificação do DNA tanto dessas bactérias como das respectivas plantas hospedeiras, sendo possível, por exemplo, o estudo do Transcriptoma, conjunto de genes expressos, tornando possível inferir sobre o envolvimento de genes responsáveis pela interação planta-bactéria.

1.3. Sequenciadores de Nova Geração

Décadas de conhecimentos, produzidos por milhares de cientistas trabalhando em genética, bioquímica, biologia celular e físico-química, produziram técnicas capazes de localizar, isolar, preparar e estudar pequenos segmentos de DNA, provenientes de cromossomos muito maiores. Esses segmentos de DNA por sua vez, são mapeados e sequenciados para conhecimento de suas bases nucleotídicas (LEHNINGER, NELSON e COX, 2006).

O método pioneiro para determinação das bases nucleotídicas foi o sequenciamento Sanger, criado na década de 70 pelo próprio Sanger e colaboradores (SANGER, NICKLEN e COULSON, 1977). A síntese de DNA inicia-se a partir de um primer marcado em uma das extremidades da fita. Dideoxinucleotídeos que não possuem grupos OH nas posições 3', bem como a posição 2' da desoxirribose são usados para terminar a síntese de DNA em bases específicas. Em seguida, quatro reações separadas são executadas simultaneamente, cada uma incluindo um desoxirrobucleotídeo (A,T,C,G) em adição ao seu homólogo natural. Assim, uma série de moléculas de DNA marcadas são geradas, representando cada uma das bases. Posteriormente, os fragmentos de DNA são separados de acordo com o seu tamanho em gel de eletroforese e expostos à raios-x para determinação das bases (COOPER, 2000). A Figura 5 ilustra todo o processo.

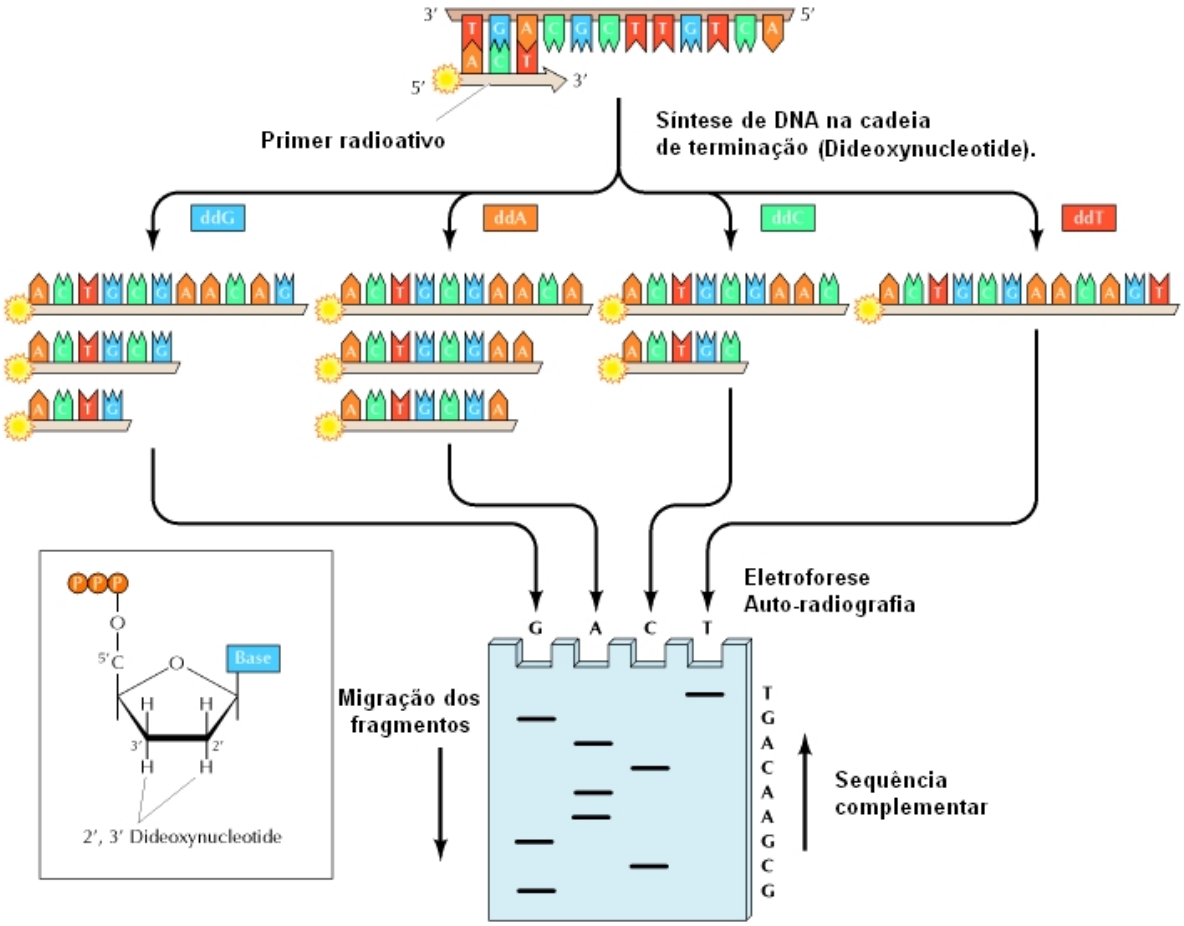


Figura 5 – Sequenciamento de DNA por meio do método de Sanger.

A síntese de DNA inicia-se com um iniciador (*primer*) radiotivo. Quatro reações separadas são realizadas, cada uma contendo um dideoxinucleotídeo misturado com o seu homólogo normal, bem como os três outros não marcados. Quando o dideoxinucleotídeo estiver incorporado, a síntese de DNA é interrompida, de modo que a reação produz uma série de produtos que se estendam a partir do *primer* da base marcada. Esses produtos das quatro reações são separadas através de eletroforese e analisadas por auto-radiografia para determinar a sequência de DNA final.

Adaptado de: (COOPER, 2000).

Ainda ressaltando as técnicas utilizadas para determinação das bases que compõem o DNA, um exemplo que deve ser citado é a estratégia de montagem do Projeto Genoma Humano (Figura 6). Clones isolados de uma biblioteca genômica foram ordenados em um mapa físico detalhado, depois clones individuais foram sequenciados por protocolos de sequenciamento aleatório de fragmentos de DNA. A estratégia utilizada pelo grupo de sequenciamento comercial eliminou a etapa de criação do mapa físico e sequenciou todo o genoma pela clonagem aleatória (VENTER, ADAMS, *et al.*, 2001).

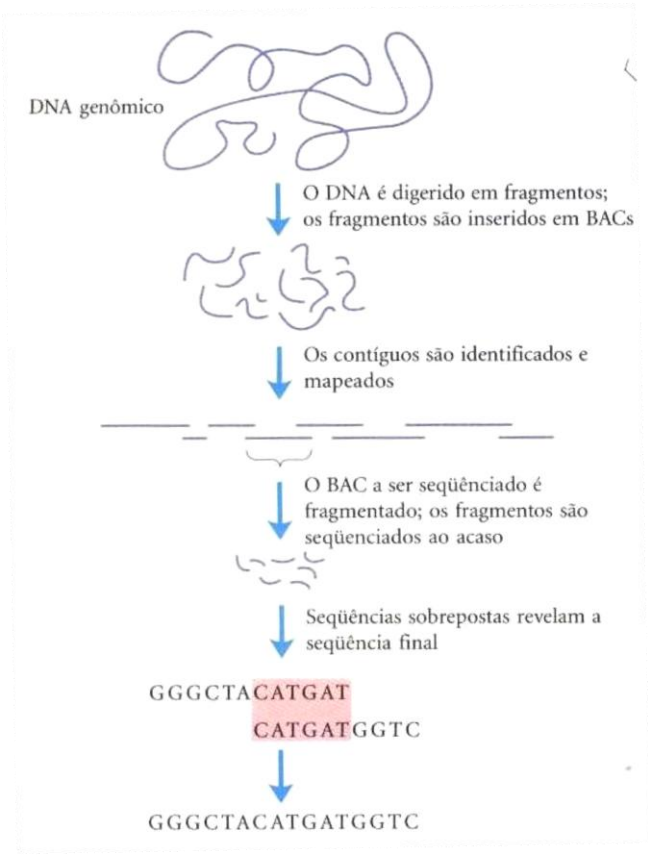


Figura 6 - A estratégia de montagem do Projeto Genoma Humano.

Inicialmente o DNA é dividido em fragmentos e incorporado aos BAC's. Os contíguos são identificados e mapeados, o BAC a ser sequenciado é fragmentado. Por fim, as sequências sobrepostas revelam a sequência final do genoma.

Fonte: (LEHNINGER, NELSON e COX, 2006).

As plataformas de sequenciamento de nova geração (NGS – *next generation sequencing*) são uma alternativa poderosa para estudos sobre genômica estrutural e funcional. Na genômica de plantas, os trabalhos com as novas plataformas têm sido destinados ao sequenciamento de transcritos, ressequenciamento ou sequenciamento *de novo* de genomas plastidiais (CARVALHO e SILVA, 2010). Este último, o genoma plastidial consiste em uma pequena molécula de DNA circular, com características semelhantes com das mitocôndrias e das bactérias. Estudos demonstram que é possível fazer análise filogenética, para diversificação de eudicotiledôneas, por exemplo, com base em genomas plastidiais (MOORE, SOLTIS, *et al.*, 2010).

Portanto, os sequenciadores NGS são uma tecnologia fundamental para o avanço das ciências biológicas modernas (LEHNINGER, NELSON e COX, 2006). Por meio dessa tecnologia é possível gerar uma enorme quantidade de dados, e com isso, aprofundar as análises sobre genomas e transcriptomas. Portanto, segue a descrição dos principais sequenciadores utilizados atualmente:

Roche 454: Foi o primeiro sequenciador a alcançar o mercado comercial, em 2004. Usa uma tecnologia de sequenciamento conhecida como pirosequenciamento. Nessa tecnologia, cada incorporação de um nucleotídeo pela enzima DNA polimerase, resulta na liberação de um pirofosfato, que inicia uma série de reações para atuação da enzima Luciferase para liberação de luz. O processo do sequenciador 454 da Roche divide-se em três etapas: a) Preparação das bibliotecas de DNA, b) PCR de Emulsão (método para amplificação de sequências DNA que utiliza uma emulsão de óleo em água para isolar as moléculas de DNA) e c) Sequenciamento. A Figura 7 ilustra o processo detalhadamente (MARDIS, 2008).

Na Figura 7 (a) O DNA é fragmentado aleatoriamente e ligado a adaptadores A/B em suas extremidades. Os fragmentos A/B são selecionados para o sequenciamento. Posteriormente, (na Figura 7b) os fragmentos são ligados às microesferas magnéticas por meio do pareamento com sequências curtas complementares presentes na superfície da microesfera. As microesferas são capturadas individualmente em gotículas oleosas onde a PCR em emulsão ocorre. Milhares de cópias do fragmento alvo são produzidas nessa fase. Por fim, (Figura 7c) as microesferas ligadas às sequências alvo fita simples são capturadas individualmente em poços no suporte de sequenciamento. São fornecidos os reagentes para a reação de pirosequenciamento, e o sinal de luz emitido é identificado a cada base incorporada, em cada poço de sequenciamento.

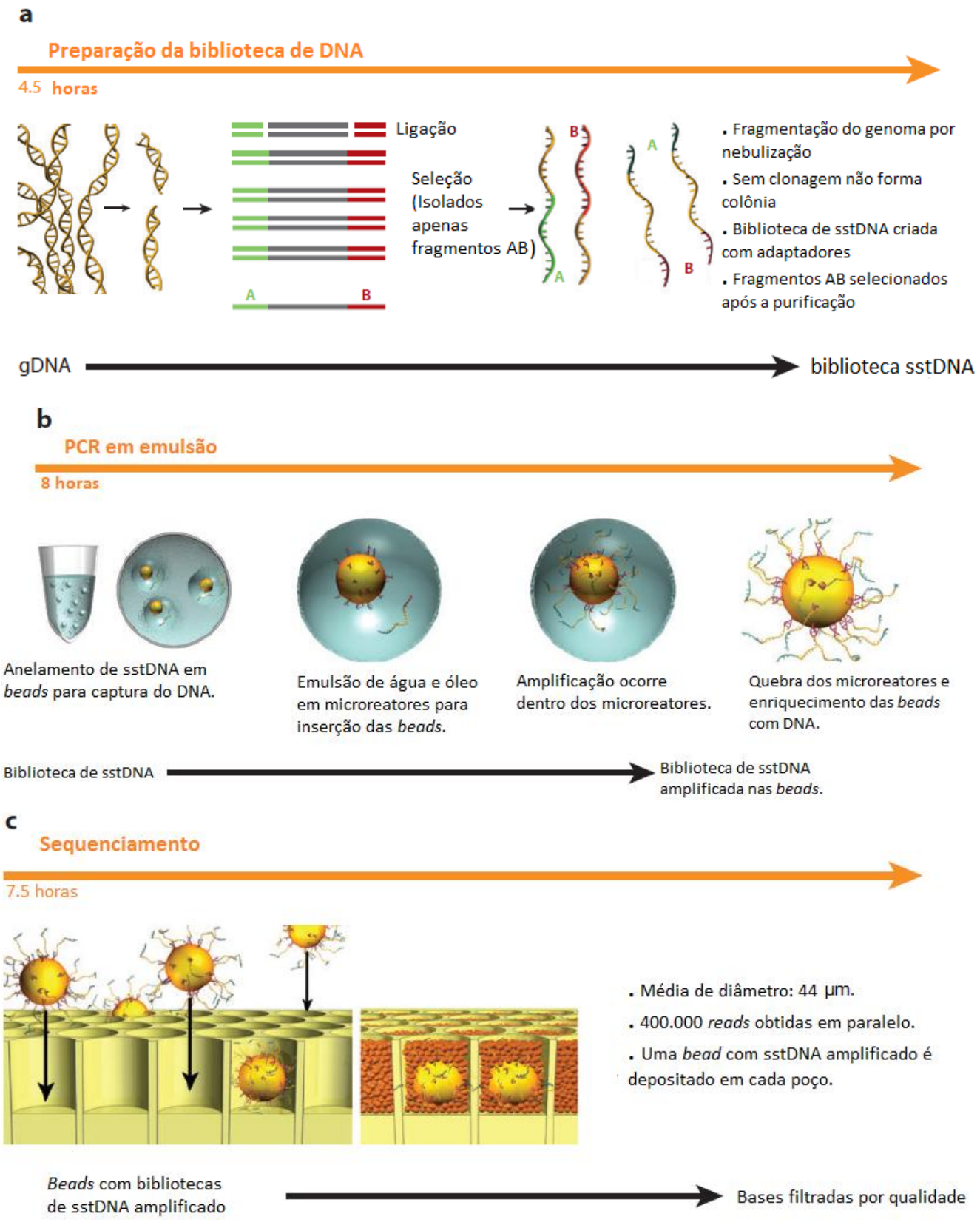


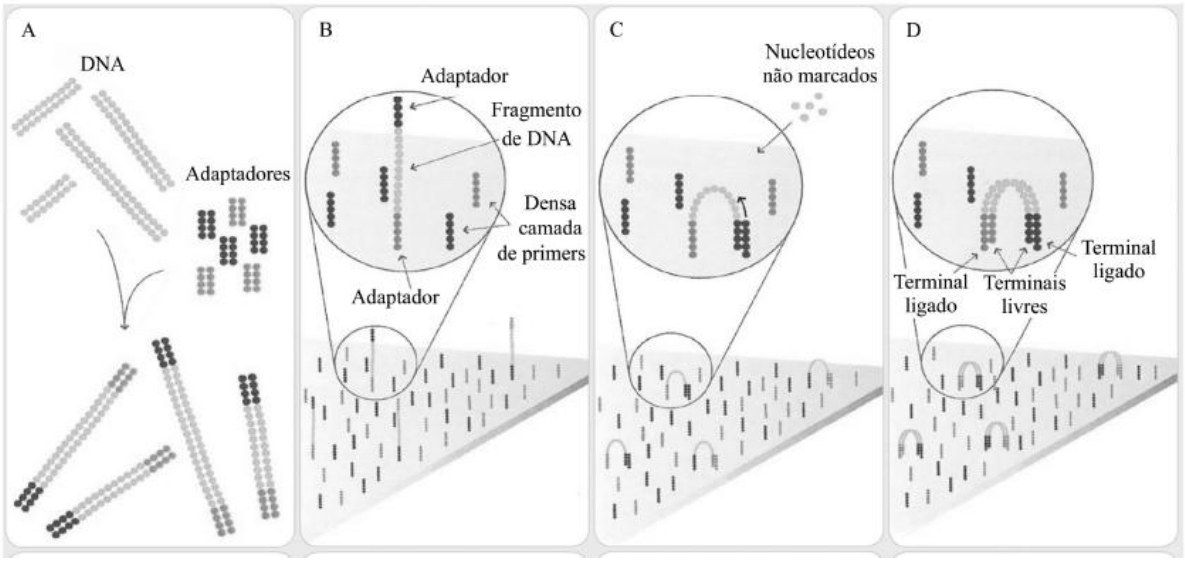
Figura 7 - Sequenciamento na plataforma 454.

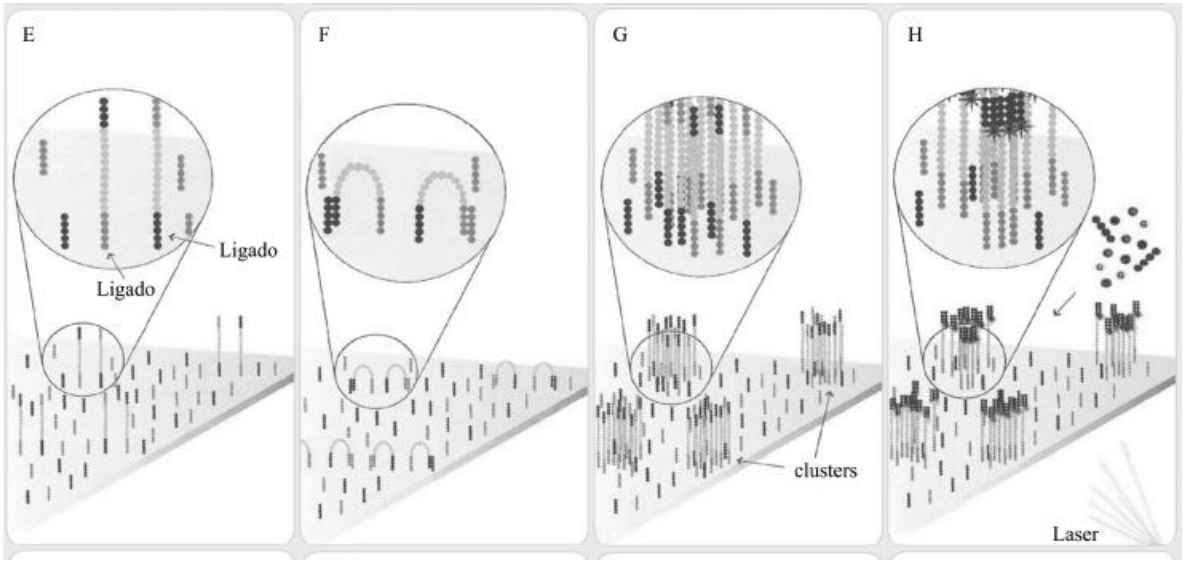
O sequenciamento é dividido em três etapas: a) Preparação da biblioteca de DNA (amostra), b) PCR em Emulsão e c) Sequenciamento. Na preparação da amostra (a), o DNA é fragmentado aleatoriamente e ligado a adaptadores. Os fragmentos A e B são selecionados. Posteriormente em (b) os fragmentos são ligados por meio de pareamento de bases presentes na superfície das *beads*. As *beads* são capturadas individualmente em gotículas de água e óleo onde ocorre a PCR em Emulsão. Por fim (c) as *beads* ligadas às sequências fita simples são capturadas individualmente no poço do suporte de sequenciamento. Em seguida, são fornecidos os reagentes para a reação de pirosequenciamento, e o sinal de luz é emitido a cada base incorporada.

Adaptado de: (MARDIS, 2008).

Illumina Solexa: As amostras de DNA recebem adaptadores específicos nas duas extremidades da molécula e os fragmentos preparados são, então, anexados à superfície de sequenciamento aleatoriamente. Em seguida, o adaptador da extremidade livre da molécula aderida ao suporte encontra seu oligonucleotídeo complementar, formando uma estrutura em ponte e os *clusters* (grupos de fragmentos) são criados (figura 8 a, b, c e d). A abordagem utilizada é chamada de *sequencing-by-synthesis* (sequenciamento por síntese – Tradução livre), onde cada nucleotídeo é marcado com uma cor (MARDIS, 2008). À medida que os nucleotídeos são incorporados, é possível identificar qual base deve ser codificada.

Na etapa de desnaturação (Figura 8 e), as pontes de hidrogênio da molécula de DNA são desfeitas mediante elevação de temperatura. Repete-se a etapa de pareamento (Figura 8 f), formando novas pontes de hidrogênio e iniciando um novo ciclo de amplificação. Após uma série de ciclos, são obtidos clusters de moléculas idênticas ligadas à superfície de sequenciamento, conforme Figura 8 g. Com a incorporação de nucleotídeos os terminadores marcados e excitação a laser (Figura 8 h), é gerado um sinal, o qual é captado e interpretado como um dos quatro possíveis nucleotídeos componentes da cadeia (Figura 8 i). O processo de incorporação dos nucleotídeos marcados, excitação e leitura é repetido para cada nucleotídeo componente da sequência nas Figuras 8 j e k. A leitura é feita de forma sequencial, o que permite a montagem da sequência completa de cada cluster (Figura 8 l) (CARVALHO e SILVA, 2010).


A
DNA
Adaptadores
B
Adaptador
Fragmento de DNA
Densa camada de primers
Adaptador
C
Nucleotídeos não marcados
D
Terminal ligado
Terminal ligado
Terminais livres


E
Ligado
Ligado
F
G
clusters
H
Laser

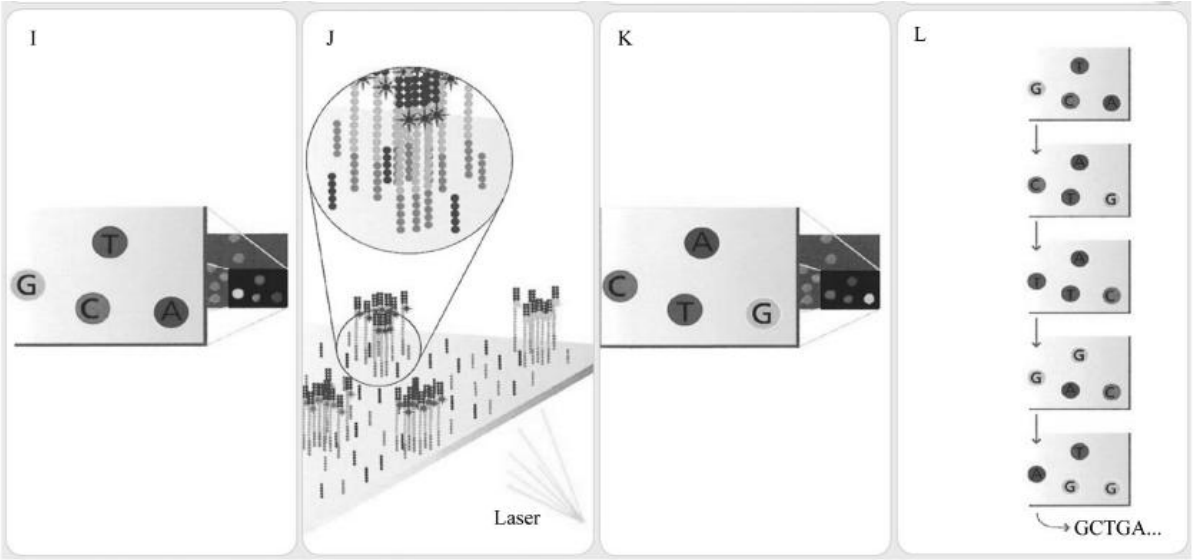


Figura 8 - Tecnologia de sequenciamento Illumina.

O DNA é fragmentado aleatoriamente e ligado a adaptadores em suas extremidades (A). As moléculas de DNA de fita simples são incorporadas ao suporte sólido por complementariedade. Durante a etapa de anelamento, no primeiro de amplificação, o adaptador da extremidade livre encontra seu oligonucleotídeo complementar no suporte, formando uma estrutura em ponte.Uma vez fornecidos os reagentes necessários, a PCR é iniciada utilizando a extremidade 3' livre como primer (B,C,D). Na etapa de desnaturação (E) a "ponte" é desfeita por elevação de temperatura. Repete-se a etapa de anelamento (F), formando novas estruturas em forma de ponte e iniciando um novo ciclo de amplificação. Após uma série de ciclos serão obtidos *clusters* de moléculas idênticas ao suporte (G). Após a incorporação de nucleotídeos terminadores marcadores e excitação a laser (H) é gerado o sinal que é captado por dispositivo de leitura e interpretado como um dos quatro nucleotídeos (I). O processo de incorporação de nucleotídeo marcado, excitação e leitura é repetido para cada nucleotídeo componente da sequência (J,K). A leitura é feita de forma sequencial, o que permite a montagem da sequência de cada *cluster* (L).

Fonte: (MARDIS, 2008)

Adaptador por: (CARVALHO e SILVA, 2010)

***Applied Biosystems* SOLiD**: A plataforma SOLiD utiliza adaptadores ligados às bibliotecas de fragmentos, semelhantes aos adaptadores de outras plataformas *NGS*, e também utiliza PCR de emulsão. Porém, o diferencial é que o SOLiD utiliza a enzima DNA ligase, e não uma polimerase. Além disso, o *output* do sequenciamento é em formato *colorspace,* ou seja, cada cor refere-se a uma dupla de nucleotídeos (dinucleotídeos) e a decodificação pode ser facilmente realizada por meio de uma tabela de cores utilizada como legenda (MARDIS, 2008). A Figura 9 apresenta os dinucleotídeos representados por cada cor. Todas as combinações possíveis de dinucleotídeos são marcadas nas sondas do SOLiD com apenas quatro fluoróforos. Assim, duas leituras de cada base são necessárias para que a sequência do dinucleotídeo seja resolvida. Esse processo inicia-se com a identificação da primeira base do alvo, ou seja, a última base do adaptador.

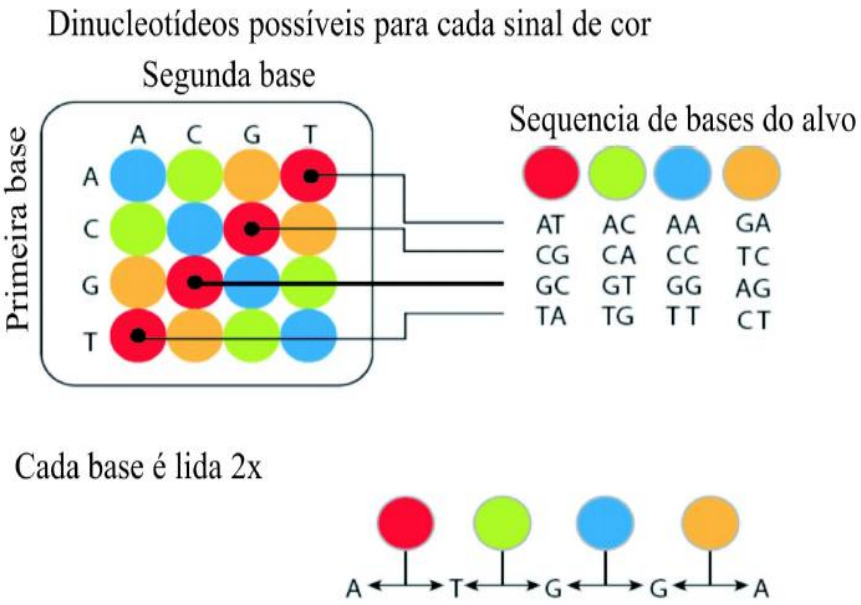


Figura 9 - Dinucleotídeos possíveis para cada sinal de cor, como saída do SOLiD.

Fonte: (MARDIS, 2008)

Adaptado por: (CARVALHO e SILVA, 2010)

A partir da identificação das cores é possível fazer a decodificação das bases, conforme apresentado na Figura 10.

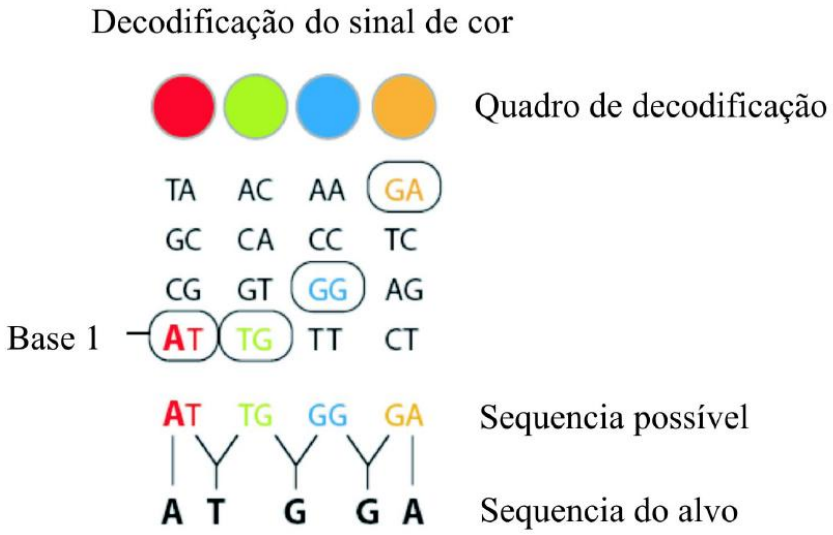


Figura 10 - Decodificação dos dinucleotídeos gerados pelo sequenciador SOLiD Biosystems.

Fonte: (MARDIS, 2008)

Adaptado por: (CARVALHO e SILVA, 2010)

Ion Torrent: Publicado em Maio de 2011, o Ion Torrent é uma tecnologia escalável com técnicas de baixo custo de fabricação. É um dispositivo de semicondutores integrados que permitem realizar sequenciamento de genomas através da detecção de níveis de pH . Semicondutores são utilizados para fazer um circuito integrado capaz de executar diretamente o sequenciamento de DNA. As sequências são obtidas por íons dirigidos pela síntese da DNA polimerase, utilizando processamento paralelo e chip iônico.

O chip contém íons sensíveis baseados em transistores, com 1.2 milhões de poços para confinamento das sequências. Além disso, permitem a detecção simultânea de reações de sequenciamento independente. Ao contrário do uso de reagentes para marcar as bases, o chip detecta uma elevação no pH que ocorre conforme cada nucleotídeo se junta à fita em crescimento e libera um próton (H^+) no processo (ROTHBERG, HINZ, *et al.*, 2011).

O Ion Torrent apresenta uma tecnologia surpreendente porque um chip é capaz de sequenciar inteiramente um genoma procariótico com uma cobertura de 20x (CUMMINGS e SCIENTIST, 2012). A Figura 11 apresenta um esquema do sensor e a arquitetura do chip. Prótons são liberados quando nucleotídeos estão incorporados nas fitas de DNA em amplificação, mudando o poço.

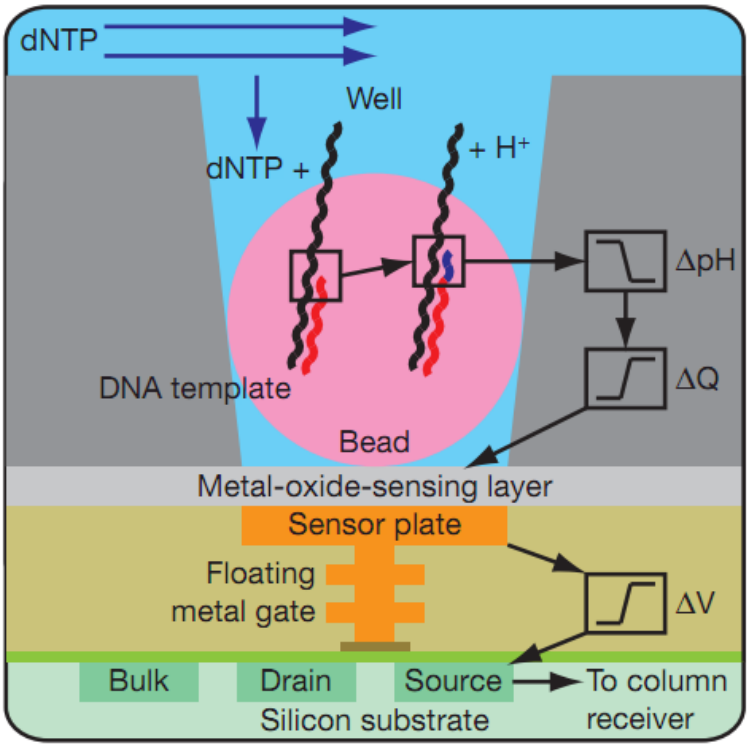


Figura 11 - Sensor e arquitetura do chip do Ion Torrent.

Funciona como um medido de pH, ou seja, é um dispositivo de semicondutores integrados que permitem realizar sequenciamento de genomas através da detecção de níveis de pH

dNTP: Nucleotídeos

H+:Prótons

ΔpH:Alteração de pH no poço

ΔV:Mudança de potencial

Fonte: (ROTHBERG, HINZ, *et al.*, 2011)

Os parágrafos anteriores descreveram algumas tecnologias de sequenciamento de nova geração (NGS), porém, a cada dia novas pesquisas são realizadas com intuito de melhorar custo, benefício e processamento de sequenciamento de DNA. Especificamente para este trabalho foram utilizadas sequências da Plataforma SOLiD3 e 4, cujo sequenciamento foi realizado por pesquisadores do Departamento de Bioquímica da Universidade Federal do Paraná (UFPR).

1.4. TRANSCRIPTOMA

O transcriptoma é o conjunto completo de transcritos da célula, em um estágio específico de desenvolvimento ou condição fisiológica. Portanto, a identificação dos transcritos expressos é essencial para o entendimento do genoma e do organismo como um todo (NOBUTA, VENU, et al., 2007). Além disso, compreender o transcriptoma é essencial para interpretação dos elementos funcionais do genoma, bem como o entendimento dos constituintes moleculares de células e tecidos ou da compreensão do desenvolvimento de doenças, por exemplo (WANG, GERSTEIN e SNYDER, 2009).

Dentro do contexto, a transcrição é um processo químico e enzimático relativamente semelhante à replicação do DNA. A principal diferença, porém, é que no caso da transcrição, a nova cadeia é formada por ribonucleotídeos e não por desoxirribonucleotídeos. Outras características que se diferem são:

- A enzima RNA-polimerase não necessita de um iniciador (*primer).* Ela sempre pode iniciar a transcrição com um promotor, ou seja, a sequência de DNA a qual a RNA-polimerase é inicialmente ligada;
- O RNA produzido não permanece ligado por pareamento de bases à fita molde de DNA. A enzima libera a cadeia em crescimento alguns nucleotídeos atrás do local em que cada ribonucleotídeo foi adicionado;
- A transcrição é menos precisa que a replicação. Ocorre um erro a cada 10.000 nucleotídeos adicionados, comparado a um erro a cada 10.000.000 da replicação.

  (WATSON, BAKER, *et al.*, 2006)

A regulação da transcrição envolve não apenas as proporções diferentes da transcrição nas diferentes partes do genoma, como também a escolha das regiões que devem ser transcritas, e a extensão desta transcrição. Desse modo, diferentes conjuntos de genes podem ser transcritos em diferentes células, ou na mesma célula, em momentos diferentes (WATSON, BAKER, *et al.*, 2006).

A transcrição passa por uma série de etapas bem definidas, sendo as principais: iniciação, alongamento e terminação. O nucleotídeo no DNA, que codifica

o início da cadeia de RNA é chamado de sítio de início da transcrição, (do inglês *Transcriptional Start Sites* – TSS), designado pela posição +1. As sequências situadas no sentido da transcrição são referidas como a jusante ao ponto de início (*downstream*). Da mesma forma, as sequências situadas na região anterior ao TSS são referidas como sequências à montante (*upstream*) (WATSON, BAKER, *et al.*, 2006).

Assim, associado a esse tema, pesquisadores da cidade de Tóquio no Japão, publicaram o artigo intitulado "*Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis*" (YAMASHITA, SATHIRA, *et al.*, 2011). Yamashita e colaboradores (2011), fizeram a análise de quatro transcriptomas diferentes, com um total de 800 milhões de *reads*[3] de 36 pb geradas pelo Illumina. Para tanto, utilizaram as técnicas de TSS-Seq, Chip-Seq, Nucleosome-Seq e RNA-Seq. Como resultado, foi possível caracterizar e *clusterizar* (agrupar) 140 milhões de TSS's encontrados em 12 tipos de células humanas. A conclusão do respectivo trabalho deixa clara a importância da integração das *OMIC's* (Genômica, Transcriptômica, Proteômica, etc.). A integração e a interpretação dos dados fornecem informações úteis para uma compreensão abrangente, neste caso, do genoma humano. No nosso caso, do arroz.

1.4.1. *Splicing* alternativo

Em organismos eucarióticos, a organização gênica consiste de sequências codificantes (éxons) separadas entre si por sequências não-codificantes (íntrons). Devido ao padrão de alternância entre éxons e íntrons, frequentemente, esses genes são denominados como "fragmentados" ou "descontínuos". O número de íntrons varia muito, desde apenas um como ocorre em genes de levedura, a 50 no gene de colágeno de galinhas e até 363, no caso do gene humano *Titin*. O tamanho de éxons e íntrons também é muito variável. Como exemplo, o gene da enzima

[3] *Reads*: termo inglês para a palavra leitura. Neste contexto, o termo mantém-se em inglês por ser amplamente utilizado pela literatura, inclusive em outras línguas.

diidrofolatorredutase de mamíferos tem mais de 31 kb de comprimento, contendo seis éxons que juntos, correspondem a 2 kb do mRNA (WATSON, BAKER, *et al.*, 2006). A Figura 12 apresenta um gene eucariótico típico no qual a região codificante é interrompida por três íntrons.

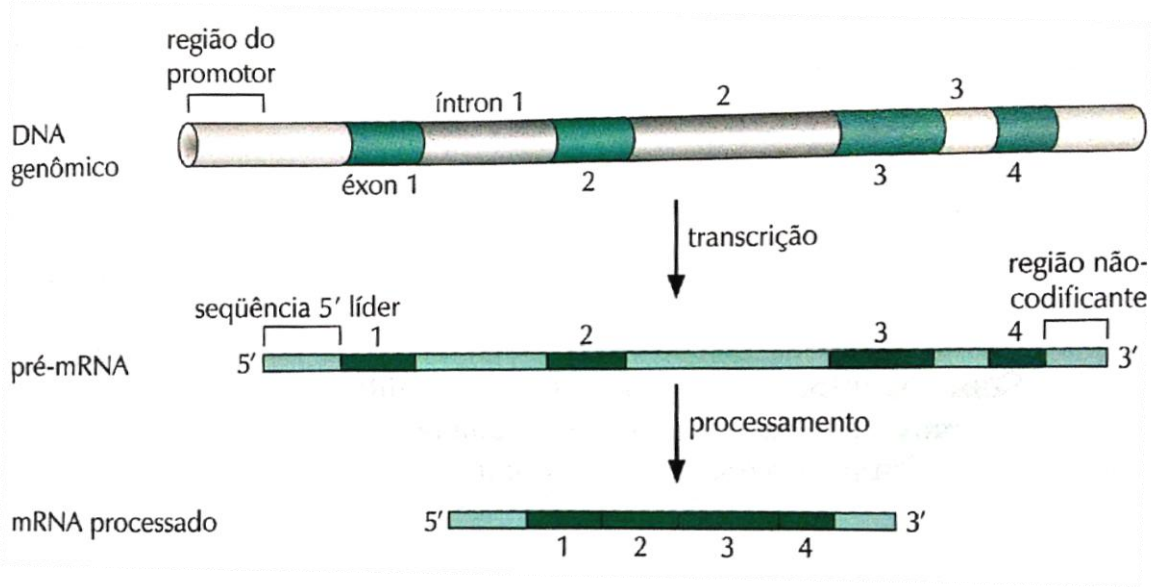


Figura 12 - Gene eucariótico.

Este apresenta quatro éxons separados por três íntrons. O processamento (*splicing*) remove os íntrons e une os éxons formando o mRNA processado.

Fonte: (WATSON, BAKER, *et al.*, 2006).

Os íntrons são removidos do pré-mRNA por meio de um mecanismo denominado processamento de RNA (*splicing*) [4]. Este processo converte o pré-mRNA em RNA mensageiro maduro. Frequentemente, pré-mRNAs podem ser processados de mais de um modo, originando mRNAs alternativos, pela remoção de diferentes combinações de íntrons. Esse processo é denominado *splicing* alternativo, dessa maneira, um gene pode dar origem a mais de um produto polipeptídico (WATSON, BAKER, *et al.*, 2006). Nota-se então que não se pode completamente afirmar que um íntron é uma sequencia não codificante, uma vez

[4] Neste contexto, o termo usado será *splicing,* por ser amplamente utilizado na literatura.

que este pode passar a compor a molécula madura de mRNA em um estágio específico da célula após um *splicing* alternativo deste gene.

O número de formas variantes que um determinado gene pode codificar pelo *splicing* alternativo varia de dois até milhares de formas. Como exemplo, o gene *Dscam* de *Drosophila melanogaster* (mosca-da-fruta). Este gene contém 95 éxons que sofrem *splicing* alternativo, ou seja, pode ser expresso em diferentes momentos ou diferentes tecidos. Este caso ilustra a enorme plasticidade do mecanismo *splicing* alternativo, onde um único gene pode expressar 38.016 produtos diferentes. O gene *Dscam* possui ainda, um total de 115 éxons abrangendo aproximadamente 60.000 pb divididos em 4 clusters (grupos) (CELOTTO e GRAVELEY, 2001). A Figura 13 mostra a organização do gene *Dscam* além do cluster denominado Exon4, destacado na figura.

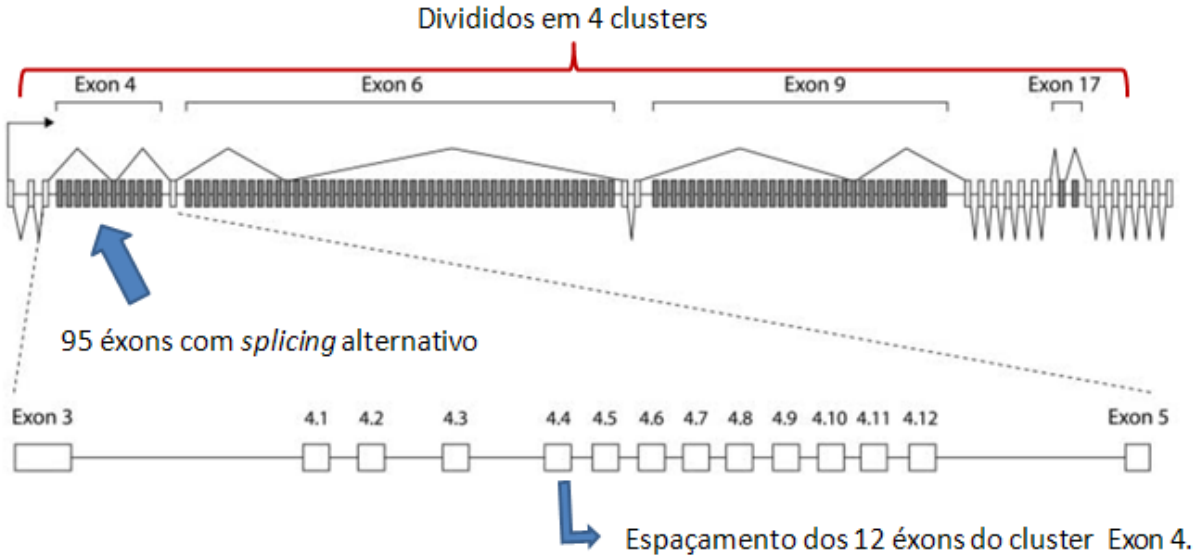


Figura 13 – Organização do gene *Dscam* de *Drosophila melanogaster*.

As caixas abertas correspondem a 20 éxons que não sofrem *splicing* alternativo. As caixas pretas correspondem aos outros 95 éxons. Os éxons de *splicing* alternativo estão organizados em 4 grupos que contêm 12, 48, 33 e 2 formas de *splicing* para cada um destes éxons. Em detalhe o cluster Exon4 para melhor visualização.

Adaptado de: (CELOTTO e GRAVELEY, 2001).

Porém, há ainda uma importante questão: Como os íntrons são retirados? Os mecanismos moleculares da reação de *splicing* precisam distinguir o que é éxon e íntron para remover os íntrons e ligar os éxons com alta precisão. Com esta

finalidade, as fronteiras entre íntrons e éxons são marcadas por sequências de nucleotídeos específicas nos pré-mRNAs. Essas sequências determinam onde ocorrerá o *splicing* e são denominadas sítio de processamento 5’ e sítio de processamento 3’. Outro fator importante é a presença de uma Adenina (Resíduo A) no sítio de ramificação (WATSON, BAKER, *et al.*, 2006). A Figura 14 apresenta as sequências consenso.

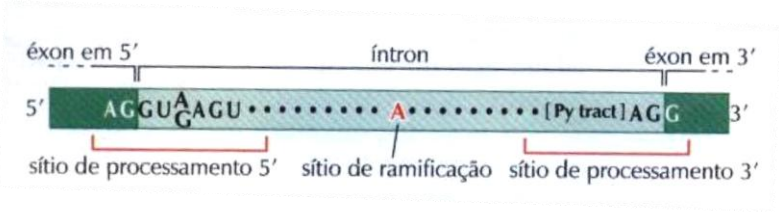


Figura 14 - Sequências na fronteira íntron-exon.

Sítio de processamento 5’ e 3’ além da Adenina no sítio de ramificação.

Fonte: (WATSON, BAKER, *et al.*, 2006).

Após o reconhecimento das regiões codificantes, ocorre a retirada do íntron, ou seja, o evento de *splicing*. Este ocorre de duas maneiras: CIS-*splicing* e TRANS-*splicing*. Na forma CIS, como acontece com o arroz, o RNA forma uma estrutura em alça. A 2’-OH da Adenina no sítio de ramificação faz um ataque nucleofílico no fosfato do resíduo G conservado no sítio de processamento 5’. Como consequência, a ligação fosfodiéster entre o açúcar e o fosfato, na junção entre íntron e éxon, é clivada e a extremidade 5’ livre do íntron é ligada ao resíduo A no sítio de ramificação. Por fim, o éxon em 5’ reverte seu papel e se torna um nucleofílico que ataca o grupo fosfato no sítio de processamento 3’, dessa maneira, os éxons 5’ e 3’ são ligados e o íntron é liberado, como grupo de saída (WATSON, BAKER, *et al.*, 2006). A Figura 15 ilustra o processo de CIS-*splicing.*

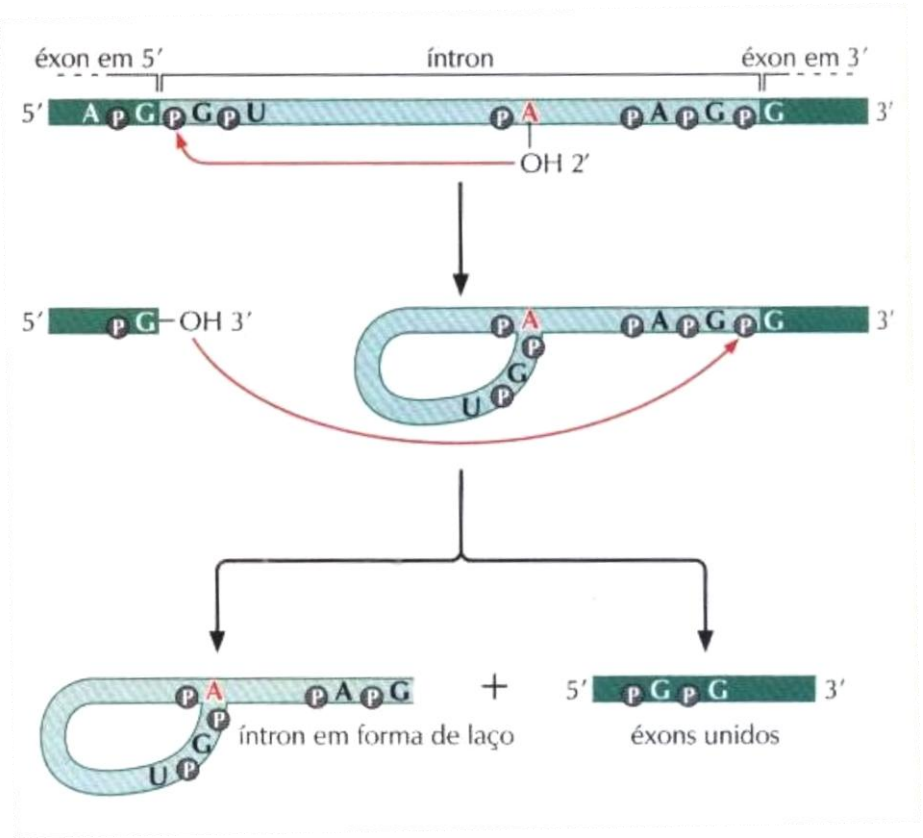


Figura 15 – *CIS-splicing*.

A 2'-OH da Adenina no sítio de ramificação faz um ataque nucleofílico no fosfato do resíduo G conservado no sítio de processamento 5'. Como consequência, a ligação fosfodiéster entre o açúcar e o fosfato, na junção entre íntron e éxon é clivada e a extremidade 5' livre do íntron é ligada ao resíduo A no sítio de ramificação (formando o laço). Na terceira etapa, o éxon em 5' reverte seu papel e se torna um nucleofílico que ataca o grupo fosfato no sítio de processamento 3', dessa maneira, os éxons 5' e 3' são ligados e o íntron é liberado, como grupo de saída.

Fonte: (WATSON, BAKER, *et al.*, 2006).

O TRANS-*splicing* tem sido encontrado principalmente em mRNAs de tripanossomos, mas outros estudos têm sido feitos para verificação desse mecanismo em outros organismos (MALEK e KNOOP, 1998). Este evento ocorre quando os éxons de diferentes transcritos primários são unidos em um novo transcrito. Ou seja, éxons localizados em diferentes moléculas de RNA podem ser ligados. Exemplo: No verme nematódeo *C. elegans*, 100% dos mRNAs sofrem TRANS-*splicing* e muitos deles também sofrem CIS-*splicing* (WATSON, BAKER, *et al.*, 2006). A Figura 16 demonstra como a reação básica de processamento é adaptada para realizar o TRANS-processamento.

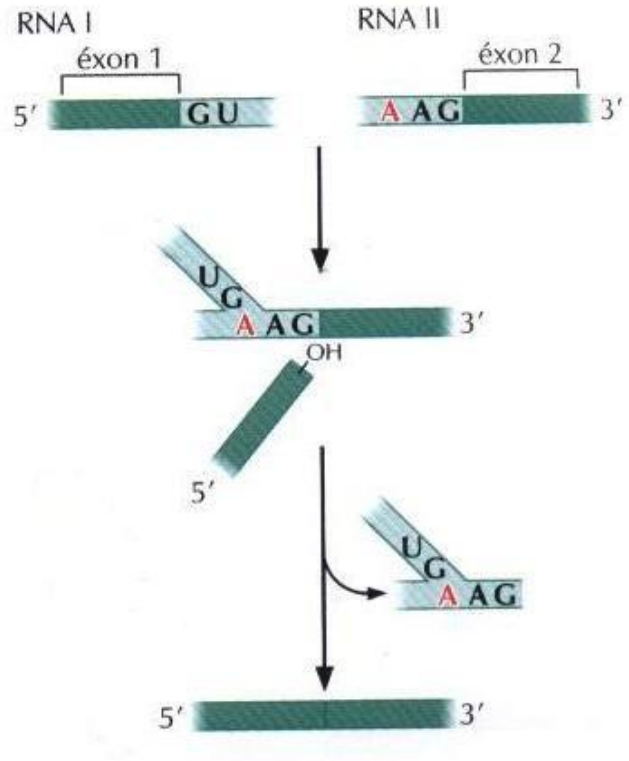


Figura 16 - TRANS-*splicing*.

A retirada dos íntrons ocorre da mesma maneira que no CIS-*splicing*, porém, éxons de diferentes transcritos primários são unidos em um novo transcrito gerando essa conformação em Y.

Fonte: (WATSON, BAKER, *et al.*, 2006).

Como é possível o genoma de *Drosophila melanogaster* conter menos genes que um genoma mais simples como de *Caenorhabditis elegans*? A resposta está no estudo do proteoma. Está se tornando claro que o *splicing* alternativo tem um papel extremamente importante na expansão da diversidade de proteínas e pode, portanto, resolver a discrepância entre o número de genes e a complexidade do organismo visto que não há proporcionalidade (GRAVELEY, 2001).

1.4.2. EST e *Microarray*

Os principais objetivos do estudo do transcriptoma são: catalogar todos os tipos de transcritos, incluindo mRNAs, RNAs não-codificantes e RNAs curtos (sRNA); determinar estrutura transcricional dos genes (em termos de seus TSS's em 5' e 3') além da verificação de *splicing* e outras modificações pós-transcricionais; e também para quantificar os níveis de expressão de cada transcrito em diferentes condições (WANG, GERSTEIN e SNYDER, 2009).

Para dedução e quantificação do transcriptoma, várias tecnologias foram desenvolvidas, dentre elas as baseadas em hibridização e as por sequenciamento. Duas técnicas importantes são: EST (*Expressed Sequence Tags*) e Microarranjo (Termo mais utilizado, do inglês: *Microarray*). Abaixo segue a descrição de cada uma delas:

EST: O método é proveniente das sequências de cDNA, onde, as etiquetas são utilizadas para identificação de transcritos bem como suas bases nucleotídicas. Este termo específico começou a ser utilizado por Adams e colaboradores em 1991, com o trabalho intitulado: "*Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project*". Porém, no fim da década de 70 pesquisadores de Harvard já trabalhavam com a idéia de fazer cópias de *mRNAs in vitro* para amplificação em uma biblioteca de dados. Além disso, trabalhos anteriores, já utilizavam bibliotecas de cDNA (SUTCLIFFE, MILNER, *et al.*, 1982) e (PUTNEY, HERLIHY e SCHIMMEL, 1983).

Primeiramente, é feita a amplificação dos genes a partir de *primers* do vetor de clonagem. Então pode-se produzir ESTs 3' ou 5', de forma que a EST 3' está ancorada ao poli-A. O conteúdo da extremidade 5', no entanto, varia devido ao tamanho do RNA, ao tamanho do inserto clonado, ao tamanho da transcriptase reversa e às diferentes formas de *splicing* do gene (PROSDOCIMI, 2012).

A Figura 17 apresenta claramente as etapas citadas. As bibliotecas de cDNA são criadas a partir da purificação de mRNA correspondente a um determinado gene, posteriormente ocorre a amplificação dos genes com utilização dos *primers,*

clonagem e sequenciamento. Após o sequenciamento serão geradas as EST's e ORESTES (EST's ricas em ORF's)

As EST's possuem aproximadamente 150 a 1000 nucleotídeos, cada clone é sequenciado apenas uma vez, permitem a identificação de *splicing* alternativo e a identificação dos genes mais expressos em diferentes fases ou tecidos.

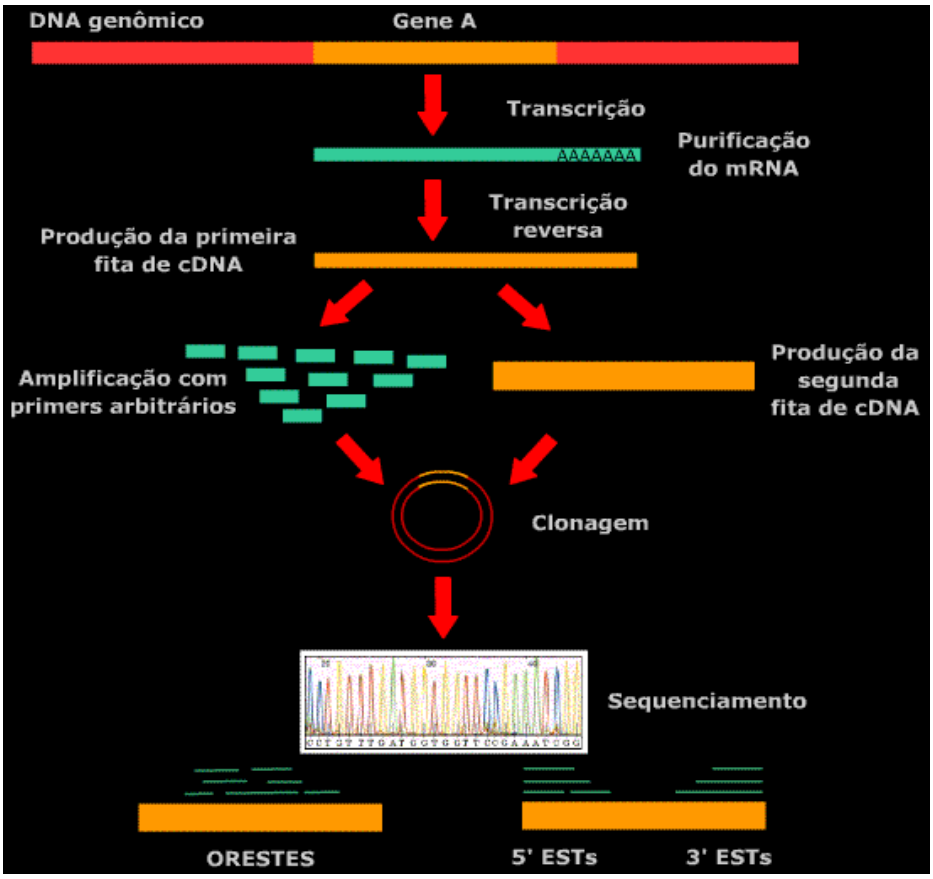


Figura 17 - Etapas de criação das EST's.

Após a criação das bibliotecas de cDNA, ocorre a amplificação das sequências com utilização dos *primers,*em seguida é realizada a etapa de clonagem e sequenciamento. Assim, são geradas as EST's e ORESTES (EST's com grande número de ORF's).

Fonte: (PROSDOCIMI, 2012).

A base de dados EST Profile do NCBI apresenta de forma visualmente interessante o nível de expressão de um determinado gene em diferentes tecidos, por exemplo. A Figura 18 apresenta o nível de expressão do gene Os08g0424500 de arroz, demonstrando em quais partes da planta o mesmo é expresso. É possível

visualizar que esse gene é expresso no caule, flor, folha, panículo, raíz e na haste de sustentação da planta.

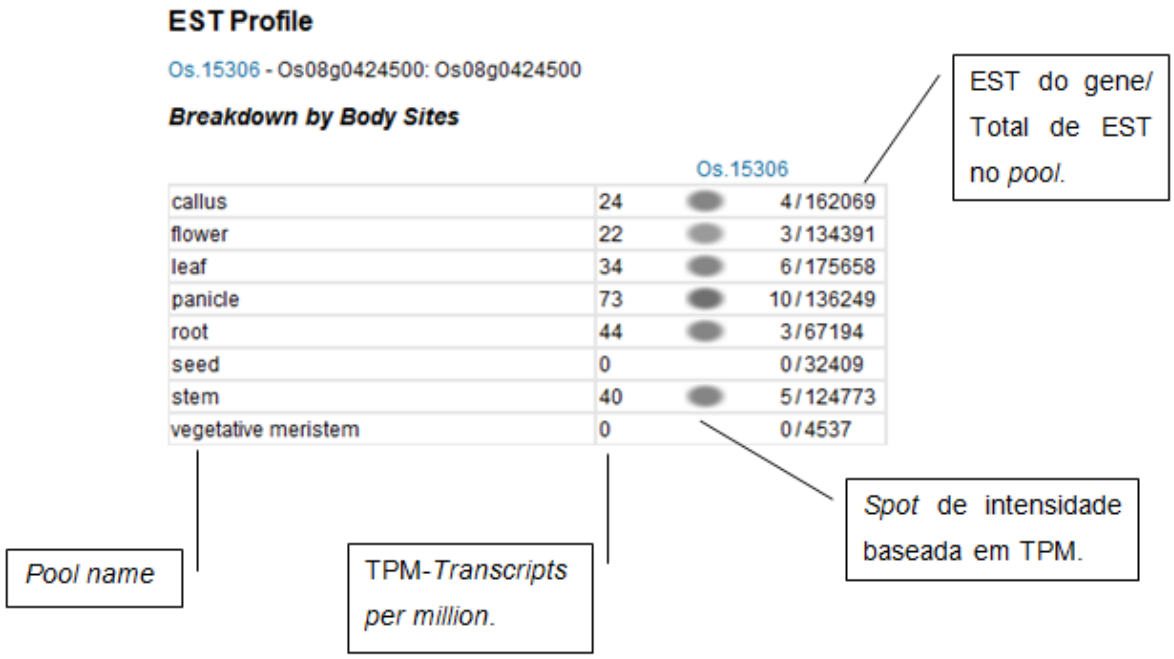


Figura 18 - Nível de expressão do gene Os08g0424500 de arroz.

Demonstra em quais partes da planta o mesmo é expresso. Ilustração criada a partir de consulta pelo nome do gene na base de dados EST Profile do banco de dados NCBI.

Microarray: Técnica baseada em hibridização para medir níveis de expressão de transcritos em escala genômica. As amostras biológicas são quimicamente ligadas a uma superfície sólida, e marcadas com fluoróforos. A matriz então é montada em sua conformação para que posteriormente seja possível fazer uma comparação entre os níveis de expressão dos genes (BRAZMA, HIGAMP, *et al.*, 2001). A Figura 19 demonstra a disposição da matriz de microarray.

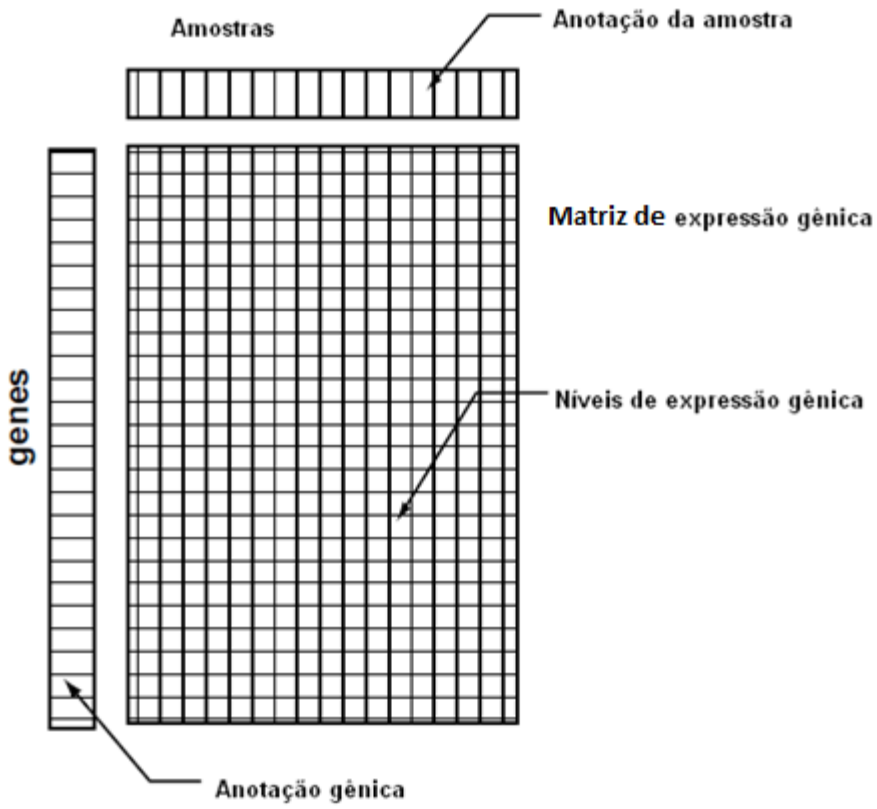


Figura 19 - Conformação da matriz de expressão gênica da técnica de microarray.

Possível verificar onde se encontram os *spots* para inserção das amostras, da anotação gênica, além da matriz com os resultados de expressão gênica.

Adaptado de: (BRAZMA, HIGAMP, *et al.*, 2001).

A Figura 20 mostra ainda, outros três níveis do processo de microarray, onde primeiro gera-se as imagens de hibridização, depois os *spots* com as matrizes de quantificação, e por último a matriz consenso com o nível de expressão gênica das amostras (BRAZMA, HIGAMP, *et al.*, 2001).

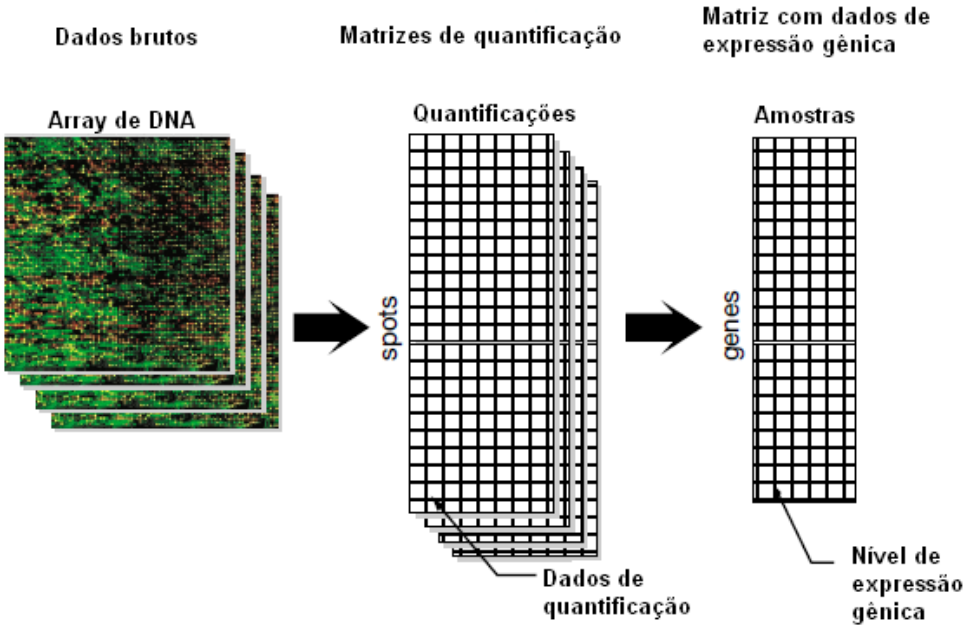


Figura 20 - Processo de microarray.

Primeiro gera-se as imagens de hibridização, depois os *spots* com as matrizes de quantificação, e por último a matriz consenso com o nível de expressão gênica das amostras

Adaptada de: (BRAZMA, HIGAMP, *et al.*, 2001).

As técnicas EST e microarray já foram muito utilizadas para análises de transcriptoma, porém, atualmente com o desenvolvimento de novas tecnologias para sequenciamento, há uma nova abordagem para mapeamento e quantificação do transcriptoma, conhecida como RNA-Seq. Este método apresenta vantagens claras sobre as abordagens existentes, como destacado na Tabela 2.

Tabela 2 - Vantagens da técnica de RNA-Seq comparada com outros métodos usados em transcriptômica.

Tecnologia	**Microarray**	**cDNA ou sequenciamento EST**	**RNA-Seq**
Especificações da tecnologia			
Princípio	Hibridização	Sequenciamento Sanger	Sequenciamento *High-throughput*
Resolução	Variável, até 100 pb	Única base	Única base
Throughput (Vazão)	Alta	Baixo	Alta
Dependência de conhecimento existente sobre a sequência genômica	Sim	Não	Em alguns casos
Ruído de fundo (*background noise*)	Alto	Baixo	Baixo
Aplicação			
Mapear regiões transcritas e expressão gênica simultaneamente	Sim	Limitada para expressão gênica	Sim
Faixa dinâmica para quantificar o nível de expressão gênica, considerando ruídos.	Uns poucos, à 100	Não se pratica	>8.000-vezes
Capacidade de distinguir diferentes isoformas	Limitado	Sim	Sim
Capacidade de distinguir expressão de alelos	Limitado	Sim	Sim
Questões práticas			
Quantidade necessária de RNA	Alta	Alta	Baixo
Custo para mapeamento de transcriptoma de grandes genomas	Alto	Alta	Relativamente baixo

Adaptado de: (WANG, GERSTEIN e SNYDER, 2009).

Conforme citado anteriormente, as sequências deste trabalho foram geradas a partir do SOLiD com uso da tecnologia de RNA-Seq, descrita no item 1.5.

1.5. RNA-Seq

Estudos atuais demonstram que o RNA-Seq têm revolucionado a maneira pela qual o transcriptoma tem sido analisado (WANG, GERSTEIN e SNYDER, 2009). O item 1.5.1 descreve o experimento de RNA-Seq e apresenta as principais vantagens observadas. Em seguida, o item 1.5.2 descreve a relevância da técnica de RNA-Seq no contexto mundial apresentando as principais pesquisas realizadas com essa nova abordagem.

1.5.1. Descrição de um típico experimento de RNA-Seq

Primeiramente, para mapear sequências curtas geradas por RNA-Seq é criada uma população de moléculas de RNA (total ou fracionada) que é convertida em uma biblioteca de fragmentos de cDNA (Figura 21).

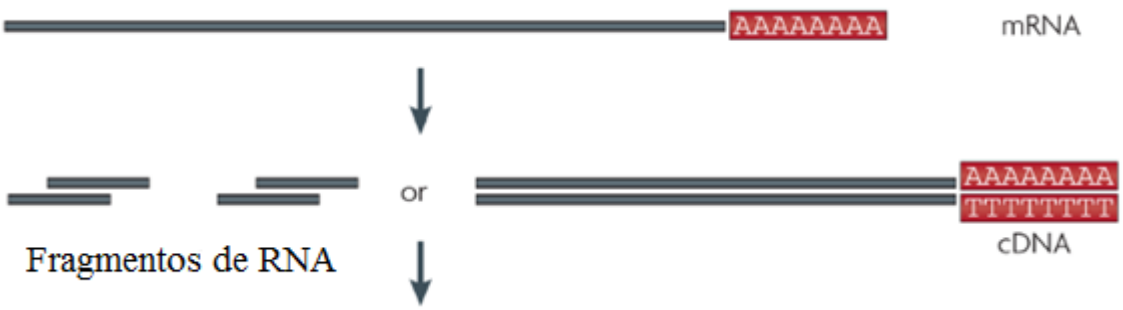


Figura 21 – Molécula de mRNA longa, convertida em bibliotecas de fragmentos de cDNA.

Fonte: (WANG, GERSTEIN e SNYDER, 2009)

Após a criação das bibliotecas de cDNA, as *reads* são geradas e alinhadas contra o genoma de referência. Posteriormente, é possível verificar as leituras que apresentam junção de éxon, e as que alinham contra um éxon das sequências do genoma de referência. A Figura 22 exemplifica este processo.

Figura 22 – Alinhamento das *reads*.

Adaptadores são adicionados a cada fragmento de cDNA e uma pequena sequência é obtida a partir das tecnologias de sequenciamento. As *reads* resultantes são então alinhadas contra o genoma de referência ou transcriptoma. Estas são classificadas em três tipos: *reads* exônicas (dentro do éxon), junções de éxons e poly(A).

Fonte: (WANG, GERSTEIN e SNYDER, 2009).

O resultado do alinhamento das *reads* contra o genoma de referência ou transcriptoma produz o nível de expressão gênica, onde, o nível de expressão do RNA é comparado com outras condições biológicas de contraste.

1.5.2. RNA-Seq e sua relevância nas pesquisas científicas

A técnica de RNA-Seq já foi aplicada a diversas análises de transcriptoma, dentre eles, ao estudo de *Saccharomyces cerevisiae, Hevea brasiliensis, Zea mays,* além de células e tecidos de mamíferos e não-mamíferos vertebrados. Portanto, segue uma revisão bibliográfica com foco em pesquisas que utilizaram RNA-Seq em diferentes organismos.

Saccharomyces cerevisiae: O principal objetivo do estudo era identificar íntrons e regiões codificadoras (ORF's). Para isso, o sequenciamento da levedura foi realizado com a tecnologia Illumina e as *reads* foram geradas com 35 pares de base (pb). Foram realizadas duas replicatas técnicas e duas replicatas biológicas. As amostras geraram um total de 15.787.335 e 14.125.182 *reads*, respectivamente.

Do total de 29.912.517 *reads*, 15.870.540 (56%) foram mapeadas em uma única região do genoma, usando *mismatch*[5] igual a 2. Como resultado, foi possível identificar os genes diferencialmente expressos e o seu nível de transcrição, além de identificar íntrons e ORF's (Figura 23) (NAGALAKSHMI, WANG, *et al.*, 2008).

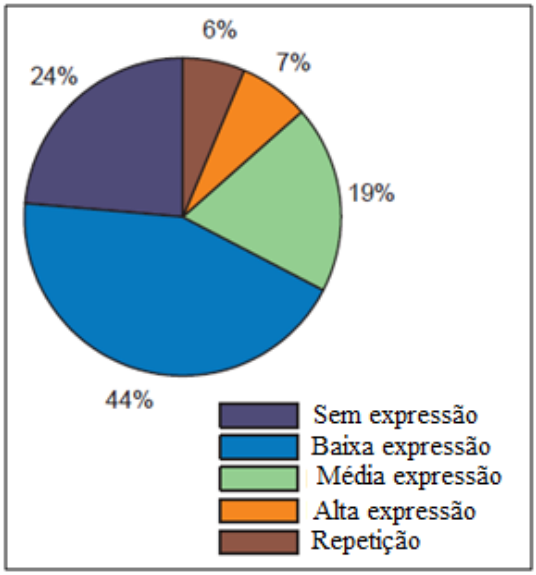


Figura 23 – Resumo do nível de expressão do organismo *Saccharomyces cerevisiae*.

Onde, "Repetição" indica o nível de sequências repetidas no genoma.

Fonte: (NAGALAKSHMI, WANG, *et al.*, 2008)

Hevea brasiliensis: A seringueira é extremamente importante economicamente por ser a única fonte de borracha natural. Assim, para facilitar a investigação biológica, bioquímica e molecular da borracha o artigo "*RNA-Seq analysis and de novo transcriptome assembly of Hevea brasiliensis*" (Xia, Xu *et al.* 2011) relata o uso de tecnologia de nova geração Illumina no sequenciamento do transcriptoma. Foram geradas mais de 12 milhões de *reads* com um tamanho médio de 90 nucleotídeos. As sequências foram anotadas com base no *Gene Ontology* (GO) e no *Clusters of Orthologous Group* (COG).

Ao todo, 37.432 genes foram anotados com sucesso, dos quais 24.545 (65,5%) foram mapeados contra sequências depositadas no NCBI e contra as

[5] *Mismatch:* Termo inglês para a condição onde não há correspondência de mapeamento.

proteínas de *Ricinus communis* (mamona). Posteriormente, os genes foram classificados de acordo com suas características biológicas e os dados gerados neste estudo fornecem o recurso de sequências mais abrangente disponível para o estudo da seringueira, bem como a análise do transcriptoma de uma espécie sem informação genômica.

Zea mays: O transcriptoma do milho foi realizado com a tecnologia de sequenciamento Illumina. Foram mapeadas mais de 120 milhões de *reads* para definir a estrutura gênica, verificar eventos de *splicing* alternativo nas células e quantificar os genes diferencialmente expressos ao longo da folha do milho. Os softwares Gbrowse e eFP foram utilizados para visualização dos dados. A Figura 24 exemplifica uma visualização do eFP *browser*. Além do estudo realizado, o conjunto de dados gerado será base para uma abordagem de biologia de sistemas para compreensão do desenvolvimento da fotossíntese (Li, Ponnala *et al.* 2010).

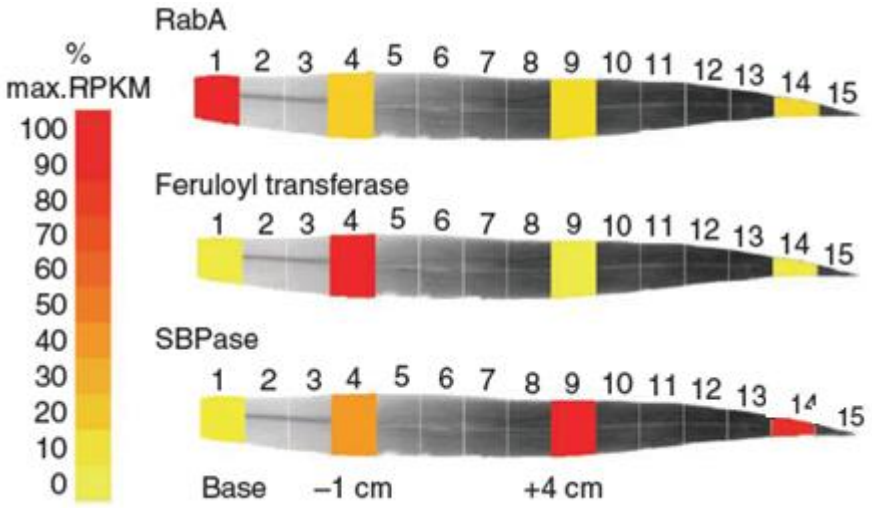


Figura 24 - Visualização da expressão gênica ao longo da folha do milho com utilização do software eFP.

Foi utilizado o cálculo de RPKM para medição dos dados de expressão ao longo da divisão em centímetros da folha do milho.

Fonte: (LI, PONNALA, *et al.*, 2010)

Células e tecidos de mamíferos: O artigo intitulado "*The evolution of gene expression levels in mammalian organs*" gerou 131 bibliotecas de cDNA sequenciadas na plataforma Illumina de 9 espécies de mamíferos: Humano,

chimpanzé, gorila, orangotango, macaco, rato, gambá e ornitorrinco. O transcriptoma gerou aproximadamente 3.2 bilhões de *reads* com 76 pb de tecidos do cérebro, cerebelo, coração, rim, fígado e testículos. Estes organismos foram escolhidos por representarem a classe de mamíferos placentários e marsupiais.

O foco deste estudo era comparar as análises de nível de expressão dos genes codificadores, e para isso, os dados foram separados em grupos de genes ortólogos e parálogos para cada espécie. Como resultado, foi possível identificar os éxons e íntrons destes genomas, verificar os genes diferencialmente expressos, gerar árvores de expressão gênica e quantificar o número de genes expressos com significância na linhagem do organismo para cada tipo de tecido estabelecido (BRAWAND, SOUMILLON, *et al.*, 2011)

Outro estudo relacionado foi realizado com vertebrados não-mamíferos. Os pesquisadores conseguiram provar que os eritrócitos não só provêem troca e transporte de gás, como podem participar ativamente da resposta do sistema imunológico. O organismo utilizado foi o *Oncorhynchus mykiss* (Truta arco-íris). As amostras foram sequenciadas no SOLiD 3, analisadas com o CLC Bio Software, as bases foram trimadas com mínimo de valor de qualidade igual a 0.05 de *score* e as sequências menores de 30 pb foram removidas. Foi utilizado o filtro de 2 *mismatches* e 5 como *match limit,* ou seja, a *read* pode alinhar até 5 vezes. (MORERA, ROHER, *et al.*, 2011).

Estes estudos demonstram que a abordagem RNA-Seq oferece arcabouço completo capaz de determinar limites de éxons, detectar e quantificar os níveis de RNAs expressos em níveis muito baixos em comparação com microarrays. Além disso, ele permite que se faça uma análise contra um genoma já sequenciado gerando até mesmo uma nova anotação dos genes (Nagalakshmi, Wang *et al*. 2008).

1.6. SOFTWARES UTILIZADOS

Os principais softwares utilizados neste trabalho foram: saet_mp – para melhorar a qualidade das *reads*, ele lê os dados e corrige baseado um conjunto de k-mers (sementes); TopHat – Mapeia as *reads* de acordo com as junções de *splicing* contra o genoma de referência; Bowtie – Utilizado pelo TopHat para alinhamento das *reads*; BEDTools – Utilizado para comparar a cobertura das *reads* no genoma de referência com o mapeamento das características (*features*) do genoma; SAMTools – Para manipulação de grandes arquivos de alinhamento das sequências de nucleotídeos; Conjunto de softwares Cufflinks – Criado como um *pipeline* de execução, ele foi utilizado para montar os transcritos, comparar os conjuntos de transcritos com a anotação do genoma e para encontrar TSS's Diferencialmente Expressos além de detectar os eventos de *splicing* alternativo.

Os itens que se segue descrevem os principais softwares utilizados.

1.6.1. Bowtie

Bowtie é um alinhador de sequências, porém, diferente do popularmente conhecido BLAST (ALTSCHUL, GISH, *et al.*, 1990). O Bowtie é otimizado para alinhar *reads* curtas em grandes genomas e foi concebido para ser extremamente rápido nesta tarefa. Considerando o genoma humano, o Bowtie alinha *reads* a uma taxa de 25.000.000 de *reads* de 35 pb por hora em uma estação de trabalho típica. É independente da plataforma utilizada e usada para gerar os índices (ou seja, a formatação dos genomas de referência), ele utiliza o algoritmo de Burrows-Wheeler, onde os dados são comprimidos. Como *output* o Bowtie gera os arquivos no formato SAM ou BAM (LANGMEAD, TRAPNELL, *et al.*, 2009). Contudo, este software trabalha com organismos procariotos porque ele não considera junções de éxon, tão comuns em eucariotos, por isso há a necessidade de se usar a ferramenta TopHat que utiliza o Bowtie internamente em seu código.

1.6.2. TopHat

TopHat é um alinhador que utiliza o Bowtie em chamadas do seu código fonte e identifica as junções éxon-éxon. Ele pode ser executado em sistemas Linux e MAC. Foi criado primeiramente para trabalhar com *reads* do sequenciador Illumina, porém, outras pessoas tem tido sucesso em utilizá-lo com outras plataformas, incluindo SOLiD, tecnologia que gera sequencias codificadas no formato *colorspace*. Como pré-requisito, o Bowtie e o SAMTools devem estar instalados no sistema. *Input*: Arquivos das bibliotecas geradas pelo SOLiD. *Output*: arquivo de alinhamento BAM, arquivo binário com o mesmo conteúdo do arquivo SAM (TRAPNELL, PACHTER e SALZBERG, 2009).

Linha de comando padrão:

tophat [opções] <índice_bowtie> <bibliotecas separadas por vírgula>

O TopHat possui uma lista variada de parâmetros, mas os principais, indispensáveis neste trabalho estão listados na Tabela 3:

Tabela 3 - Parâmetros do TopHat indispensáveis para organismos eucariotos e sequenciador SOLiD.

Fonte: (TRAPNELL, PACHTER e SALZBERG, 2009).

Parâmetro	Função
- N	Número de *mismatches* permitidos. Neste trabalho foi utilizado o default = 2.
--bowtie1	Porque o Bowtie 2.0 não suporta o formato Colorspace.
-o/--output-dir	Diretório de saída dos dados.
-I/--max-intron-length	Tamanho máximo do íntron. O default é 500.000 pb, aqui foi utilizado o valor 10.000 conforme literatura.
-C/--color	Formato Colorspace do SOLiD.
--library-type	Escolher o tipo de sequenciador utilizado. Utilizado = fr-secondstrand por serem *reads* do SOLiD.
-G/--GTF	Arquivo de anotação para ancoragem do alinhamento.
-p	Número de processadores. Foi utilizado o valor 20.
genome	Arquivos de índice gerados pelo Bowtie-build.

1.6.3. Cufflinks

Com os arquivos de alinhamento no formato SAM ou BAM, gerados pelo Bowtie ou TopHat, é possível utilizar o *pipeline* Cufflinks. Este usa os seguintes pacotes: Cufflinks para montagem dos transcritos; Cuffcompare para comparar os transcritos montados com a anotação; Cuffmerge que junta as montagens dos transcritos e gera os arquivos de anotação dos mesmos (.gtf); e Cuffdiff para encontrar genes e TSS's diferencialmente expressos, além de detectar os *splicing's* (TRAPNELL, ROBERTS, *et al.*, 2012). A Figura 25 ilustra todo o processo realizado pelo *pipeline* Cufflinks.

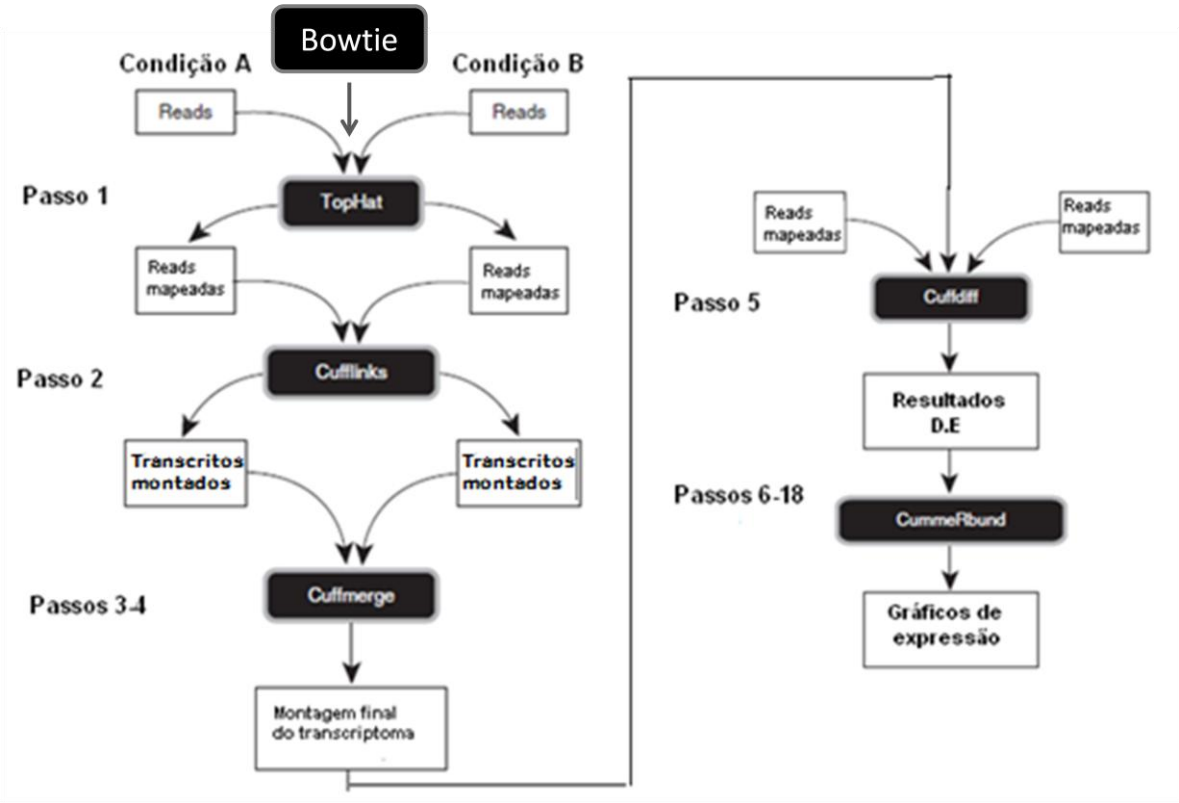


Figura 25 - Etapas do *pipeline* de execução Cufflinks.

O software Bowtie cria os índices do genoma necessários como *input* do software TopHat. O TopHat faz o mapeamento das *reads* do transcriptoma contra o genoma de referência. Em seguinda, o *pipeline* Cufflinks monta os transcritos a partir dos arquivos de alinhamento gerados pelo TopHat (Passos 2-4). Para análise dos resultados D.E (*Differential Expression*) é utilizado o módulo Cuffdiff. Por último, para geração dos gráficos é utilizada a biblioteca CummeRbund do R Bioconductor.

Adaptado de: (TRAPNELL, ROBERTS, *et al.*, 2012).

O Cufflinks utiliza o cálculo de FPKM (*Fragments per kilobase of exon per million fragments mapped*), que utiliza o conceito de RPKM (*Reads per kilobase per million mapped reads*) para normalização dos dados, visto que, transcritos maiores geram mais *reads* que transcritos menores. Porém, ele considera uma contagem por fragmentos e não por *reads*. Quando um gene sofre *splicing* alternativo e produz múltiplas isoformas na mesma amostra, muitas *reads* irão mapear em éxons compartilhados, o que complica o processo de contagem das *reads* para cada transcrito. Assim, para calcular com precisão o nível de expressão para cada transcrito, um procedimento simples de contagem não é suficiente. Portanto, Cufflinks e Cuffdiff implementam um modelo estatístico linear que observa as *reads*

com maior semelhança (método estatístico conhecido como *maximum likelihood*) (TRAPNELL, ROBERTS, *et al.*, 2012).

Cálculo de RPKM:

$$RPKM = \frac{10^9 * C}{N * L}$$

Onde:

C = Total de *reads*

N = *Reads* da amostra

L = Tamanho da transcrição

10^9 = 1000 *pb* * 1.000.000 de *reads*

Fonte: (MORTAZAVI, WILLIANS, *et al.*, 2008).

Neste trabalho foi utilizada a versão 1.2.1 por ser a versão descrita no artigo: "*Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks*." Nature Protocols (TRAPNELL, ROBERTS, *et al.*, 2012). Este artigo descreve os experimentos executados e um protocolo de execução do programa.

Abaixo, seguem as linhas de comando padrão dos módulos do Cufflinks, bem como os arquivos de entrada e saída (Input/ Output):

Cufflinks:

Linha de comando: cufflinks [opções] <arquivos de alinhamento (sam/bam)

Input: Arquivos de alinhamento (sam/bam);

Output: Para cada biblioteca é gerada uma pasta contendo os seguintes arquivos:

genes.fpkm_tracking = Valor de FPKM para os genes.

isoforms.fpkm_tracking = Valor de FPKM para as isoformas.

skipped.gtf = Arquivo de transcritos não utilizados. Neste trabalho este arquivo ficou vazio para todas as bibliotecas.

transcripts.gtf = Transcritos montados de uma determinada biblioteca.

Cuffmerge:

Linha de comando:
cuffmerge [opções] <Arquivo .txt com o caminho dos transcritos montados>

Input: Arquivo texto (.txt) contendo o caminho dos arquivos de montagem dos transcritos de acordo com o experimento. Por exemplo, se vão ser comparados os dados de controle e inoculado (CR, IR) das bibliotecas de 3 dias do cromossomo 10, o arquivo contém as linhas demonstradas na Figura 26:

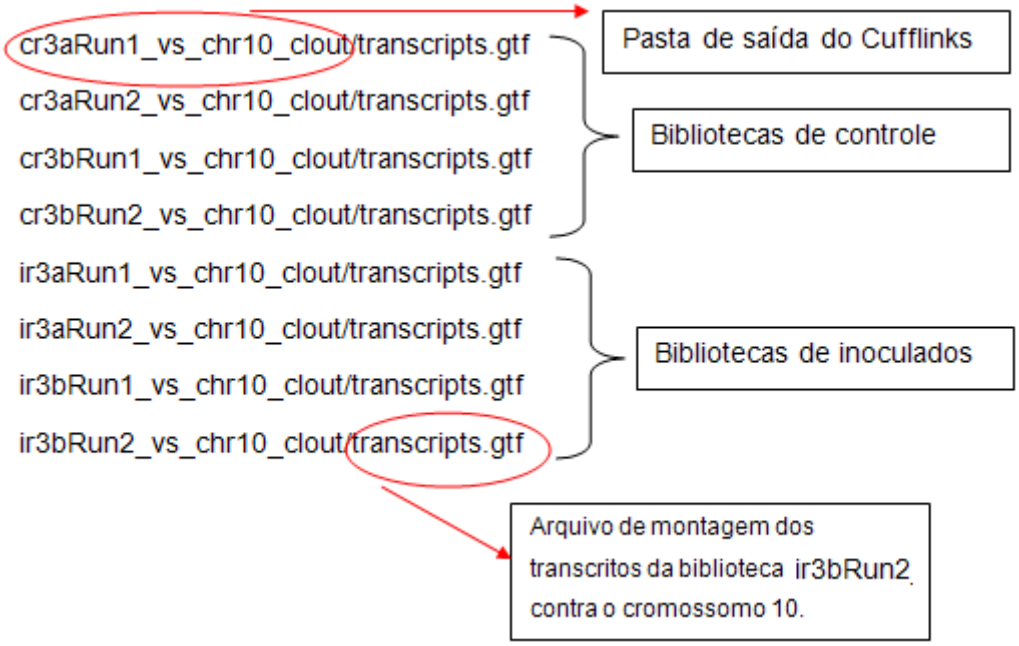


Figura 26 - Arquivo para montagem dos transcritos.

Este contém, o caminho dos arquivos .gtf das bibliotecas de controle e inoculados.

Output: Pasta com os mesmos arquivos gerados pelo Cufflinks, porém, não mais por biblioteca, mas agora com a montagem de todos os transcritos conforme especificado no arquivo assemblies.txt, por exemplo (Demonstrado acima).

Cuffdiff:

Linha de comando:

cuffdiff [opções] <transcripts.gtf>
amostra1_replicata1.sam[,...,amostra1_replicataN]>
<amostra2_replicata1.sam[,...,amostra2_replicataN.sam]>...
[amostraN.sam_replicata1.sam[,...,amostra2_replicataN.sam]]

Input: Arquivos de montagem gerados pelo Cuffmerge (.gtf) e os arquivos de alinhamento do TopHat (BAM ou SAM).

Output: Os arquivos .fpkm_tracking contêm as contagens e os valores de FPKM para os CDS, genes, isoforms e TSS. São eles: cds.fpkm_tracking, genes.fpkm_tracking, isoforms.fpkm_tracking,tss_groups.fpkm_tracking.

Os arquivos .diff contêm os dados de contagem, mais a informação na última coluna informando se o item é Diferencialmente Expresso (D.E) ou não. São eles: splicing.diff, cds_exp.diff, gene_exp.diff, isoform_exp.diff, promoters.diff, tss_group_exp.diff, cds.diff.

A lista completa de parâmetros do Cufflinks está disponível em <http://cufflinks.cbcb.umd.edu/manual.html#cufflinks> (Acesso em 25 de outubro de 2012). No entanto, os principais parâmetros utilizados estão listados na Tabela 4:

Tabela 4 - Principais parâmetros utilizados pelo Cufflinks.

Cufflinks	
Parâmetro	Função
-o/--output-dir	Diretório de armazenamento dos dados.
-p/--num-th*reads*	Número de processadores.
Cuffmerge	
-o <outprefix>	Nome da pasta de saída.
-g/--ref-gtf	Uso do arquivo de anotação (.gff).
-p/--num-th*reads*	Número de processadores.
-s/--ref-sequence	Arquivo de genoma (.fasta).
Cuffdiff	
-o/--output-dir	Diretório de armazenamento dos dados.
-L/--labels	Nome das amostras. Foi utilizado: CR3 (Controle 3 dias) CR7 (Controle 7 dias) IR3 (Inoculado 3 dias) IR7 (Inoculado 7 dias)
-p/--num-th*reads*	Número de processadores.
-u/--multi-read-correct	Estimação inicial de mapeamento dos transcritos em vários locais do genoma. Aqui é passado o arquivo de montagem dos trancritos (.gtf) gerado pelo Cuffmerge.

Após gerar os arquivos .diff através do programa Cuffdiff, o pacote CummeRbund do R/Bioconductor foi usado para criar os gráficos relacionados à análise.

1.6.4. R Bioconductor e CummeRbund

Bioconductor fornece ferramentas para análise e compreensão de dados de sequenciamento *high-throughput,* como os gerados pelos Sequenciadores de Nova Geração. Utiliza a linguagem de programação para estatística “R” e é *open source* (código aberto – Tradução livre). Disponível em: <http://www.bioconductor.org/> (Acesso em 25 de outubro de 2012).

O CummeRbund é um pacote do R/Bioconductor projetado para simplificar a visualização de dados de RNA-Seq. Diversos tipos de gráficos são gerados, dentre

eles: Densidade, Box plot, Dispersão e Dendograma. Após se conectar no R, carregar a biblioteca CummeRbund, instancia-se uma variável (cuff, por exemplo) com os dados de contagem e em seguida as linhas de comando são muito simples:

- Densidade gênica: csDensity(genes(cuff))

  Eixo x: Valor de normalização do tamanho do transcrito em log10(fpkm).

  Eixo y: Densidade, ou seja, a distribuição do nível de expressão para cada amostra.

- Box plot: csBoxplot(genes(cuff),replicates=T)

  Eixo x: Valores de log10(fpkm)

  Eixo y: Nome da amostra, Exemplo: CR3 e IR3.

- Dispersão: csScatter(genes(cuff),"hESC","Fibroblasts",smooth=T); Onde “hESC","Fibroblasts" são as amostras, que no nosso caso são (CR,IR). Os eixos x e y apresentam os valores de expressão de uma amostra contra a outra, em FPKM, então os eixos podem ficar da seguinte maneira:

  Eixo x: Nome da amostra, que neste caso é IR3.

  Eixo y: Nome da amostra, que neste caso é CR3.

- Dendograma: csDendro(genes(cuff)); Este gráfico é utilizado para visualização de *clusters* de genes, por exemplo. Por *default*, é utilizada a métrica *Jensen-Shannon* que mede a similaridade entre duas condições e calcula a distância entre as amostras.

  Eixo x: Nome das amostras comparadas. Como teste foram utilizadas as replicatas IR3b, CR3b, CR3a, IR3a.

  Eixo y: Valores de distância para agrupamento.

  (TRAPNELL, ROBERTS, *et al.*, 2012).

Os scripts para execução desses softwares e outros criados para alguma necessidade própria, foram gerados principalmente nas linguagens Python e Shell e serão mencionados no item de Materiais e Métodos.

2. OBJETIVOS

2.1. Objetivo geral

A literatura indica que o *splicing* alternativo é sem dúvida, um dos mecanismos que confere complexidade ao organismo, visto que a complexidade não deve ser comparada ao número de genes, por exemplo. Assim, esse trabalho visa identificar *splicings* alternativos em sítios de início de transcrição (TSS) de arroz utilizando dados de RNA-Seq.

2.2. Objetivos específicos

- Verificar TSS's diferencialmente expressos;
- Identificar transcritos novos;
- Caracterização do mapeamento das *reads* e dos genes expressos.

3. MATERIAIS E MÉTODOS

3.1. Obtenção dos dados

Para este trabalho foram utilizadas *reads* provenientes do sequenciador SOLiD 3 e 4. O sequenciamento, criação do cDNA e o *trimming* (corte de bases) foi realizado pelo Núcleo de Fixação Biológica de Nitrogênio no Departamento de Bioquímica, Biologia Molecular e Genética da Universidade Federal do Paraná pela aluna Liziane Cristina Campos Brusamarello Santos durante seu Doutorado (Dados não publicados). O experimento foi realizado da seguinte maneira:

A planta de arroz (*Oryza sativa* subsp. *japonica*) foi inoculada com a bactéria fixadora de nitrogênio, *Herbaspirillum seropedicae* estirpe SmR1. Em seguida o RNA foi extraído da raiz do arroz, em quatro grupos: inoculado e controle após 3 e 7 dias de inoculação, respectivamente. Foram criadas, então, 8 bibliotecas de cDNA, das quais 4 bibliotecas referentes a 3 dias após a inoculação e 4 bibliotecas de 7 dias. Porém, como as bibliotecas de 3 dias foram sequenciadas 2 vezes, então computacionalmente falando, foram geradas 12 bibliotecas de cDNA, ou seja 12 arquivos, dos quais 8 são referentes a 3 dias após a inoculação e 4 a 7 dias.

Foram feitas replicatas técnicas e biológicas, a Tabela 5 apresenta o nome de cada biblioteca e a contagem das *reads*. Onde, “cr" significa controle, “ir” inoculado, “a” e “b” são replicatas biológicas, Run1 indica a primeira “corrida” no sequenciador Solid3 e Run2 a segunda “corrida”, esta realizada no Solid4. As *reads* foram trimadas no software CLCBio, com *Score* mínimo de qualidade de 0,05 e *Phred* igual a 13. Em vermelho, o total de *reads* utilizadas: 328.409.635.

Tabela 5 - Contagem das *reads* de arroz.

Arquivo	Brutas	Trimadas	Utilização
cr3aRun1.fa	10.411.547	7.249.398	69,63%
cr3aRun2.fa	27.220.318	17.317.465	63,62%
cr3bRun1.fa	16.710.422	11.776.818	70,48%
cr3bRun2.fa	47.410.453	28.755.639	60,65%
cr7aRun1.fa	141.889.184	91.009.317	64,14%
cr7bRun1.fa	74.178.948	41.411.264	55,83%
ir3aRun1.fa	17.137.481	12.441.077	72,60%
ir3aRun2.fa	34.460.699	20.557.530	59,66%
ir3bRun1.fa	8.083.021	5.465.191	67,61%
ir3bRun2.fa	36.964.475	18.307.472	49,53%
ir7aRun1.fa	60.944.261	41.773.493	68,54%
ir7bRun1.fa	51.384.452	32.344.971	62,95%
TOTAL	526.795.261	328.409.635	62,34%
EXCLUÍDAS		198.385.626	

Como genoma de referência para alinhamento e montagem dos transcritos, foram utilizadas a versão 5 de anotação e sequências do site RAP-DB, disponível em: < http://rapdb.dna.affrc.go.jp/>. Para representação da mitocôndria e do cloroplasto os dados foram retirados no site NCBI e gerado um arquivo de anotação para cada.

3.2. CONFIGURAÇÃO DE SISTEMAS

3.2.1. Clusters computacionais e servidores

Para análise dos dados foram utilizados dois clusters computacionais:

- Cluster da Bioinformática (UFPR), com 512 GB de memória RAM, 64 processadores e 2 máquinas virtuais.
- Cluster da UFRGS (Denominado Gauss), com 64 unidades de processamento, cada qual com 64 GB de memória RAM e 2 processadores dodecacore. Um total de 1536 núcleos de processamento. A Figura 27 demonstra o cluster da SGI-Altix, Gauss.

Figura 27 - Cluster SGI-Altix, Gauss da UFRGS.

Fonte: http://www.cesup.ufrgs.br/modules/cesup/?s=hardware

Além dos clusters foram utilizados um servidor de banco de dados, e um computador local localizados no Laboratório de Bioinformática da UFPR.

- Servidor de banco de dados MySQL com 265 GB de espaço em disco, 16 GB de RAM e Sistema Operacional Linux Ubuntu.
- Computador local com 160 GB de espaço em disco, 16 GB de RAM, Dual Boot com os Sistemas Operacionais Windows7 e Linux Ubuntu. Foi necessário a instalação dos pacotes Python, R/Bioconductor e CummeRbund.

O mapeamento do transcriptoma do arroz contra o genoma de referência realizado no Supercomputador Gauss utilizando o Software TopHat, obteve um tempo de execução médio de aproximadamente uma semana sem interrupções. As análises do Cufflinks tanto no Gauss, quanto no cluster da Bioinformática demoram mais 3 dias. Finalmente, a etapa de finalização e criação dos gráficos é executada com relativa rapidez, sendo esta realizada no computador local. Esta etapa dura, em média, um dia adicional após a definição dos gráficos necessários. A saída do TopHat somada à saída do Cufflinks e as bibliotecas gerou uma quantidade de dados com cerca de 35 GB de espaço em disco. Porém, a análise total com exceção dos gráficos e incluindo os testes criou aproximadamente 507 GB de dados.

3.2.2. Etapas da parte computacional e scripts utilizados

1ª etapa: Criação do índice (formatação do genoma de referência) utilizando o software Bowtie. A Figura 28 demonstra o script desenvolvido em Shell e destaca o comando Bowtie-build utilizado para gerar o índice, denominado neste caso como rice_chr09e10.

bowtie_9e10.sh

```
#! /bin/bash

#PBS -S /bin/sh
#PBS -N bowtie_9e10
#PBS -j oe
#PBS -l select=1:ncpus=24
#PBS -m ae
#PBS -M katiaplopes@gmail.com

cd $PBS_O_WORKDIR

#Executar do caminho: /home/u/abarbosa/home/Documentos/testes_cufflinks/chr09e10

export PATH=/home/u/abarbosa/bin/:$PATH

genoma="/home/u/abarbosa/home/Documentos/oryza_sativa/genoma"

bowtie-build -f $genoma/osat_chr09.fasta,$genoma/osat_chr10.fasta rice_chr09e10
```

Figura 28 - Criação do índice do genoma de arroz com o software Bowtie.

O comando bowtie-build é utilizado para criação do mesmo, juntamente com os parâmetros de caminho dos arquivos fasta.

2ª etapa: Mapeamento do transcriptoma contra o genoma de referência. Dados informacionais:

- Genoma de referência: RAP-DB
- 1 biblioteca x cada cromossomo
- 168 arquivos .BAM
- 1 semana de processamento
- Bases trimadas
- Mismatch 2

A Figura 29 demonstra o script criado e destaca os parâmetros indispensáveis utilizados para mapeamento de organismo eucarioto.

Figura 29 - Alinhamento com o software TopHat.

Foram utilizados os parâmetros de número de processadores, tamanho máximo de íntron, mismatch, além dos caminhos dos arquivos de anotação e das bibliotecas de cDNA.

Após o mapeamento foram gerados 168 arquivos BAM que é a representação binária do alinhamento das *reads* contra genoma de referência representado no formato SAM (12 bibliotecas x cada cromossomo, incluindo as organelas = 168 arquivos BAM). Este possui os campos listados abaixo (A especificação total pode ser lida em *The Sam Format Specification* disponível em < http://samtools.sourceforge.net/SAM1.pdf> Acesso em 03 de Nov. de 2012).

Arquivo .sam = .bam

- 1: Query name
- 2: Bitwise flag* (Cada situação um valor diferente)
- 3: Reference sequence name
- 4: Left most position of mapping
- 5: Mapping quality
- 6: CIGAR

- 7: Name of mate
- 8: Position of mate
- 9: Observed template length
- 10: Sequence
- 11: Qual.: ASCII of Phred-scaled base QUALity+33[6]

Outro arquivo que deve ser utilizado é o arquivo de anotação, com extensão gff. É um arquivo texto, contínuo, conforme demonstra a Figura 30. A seta indica que as linhas continuam sem quebra e os nomes dos campos foram mantidos como são.

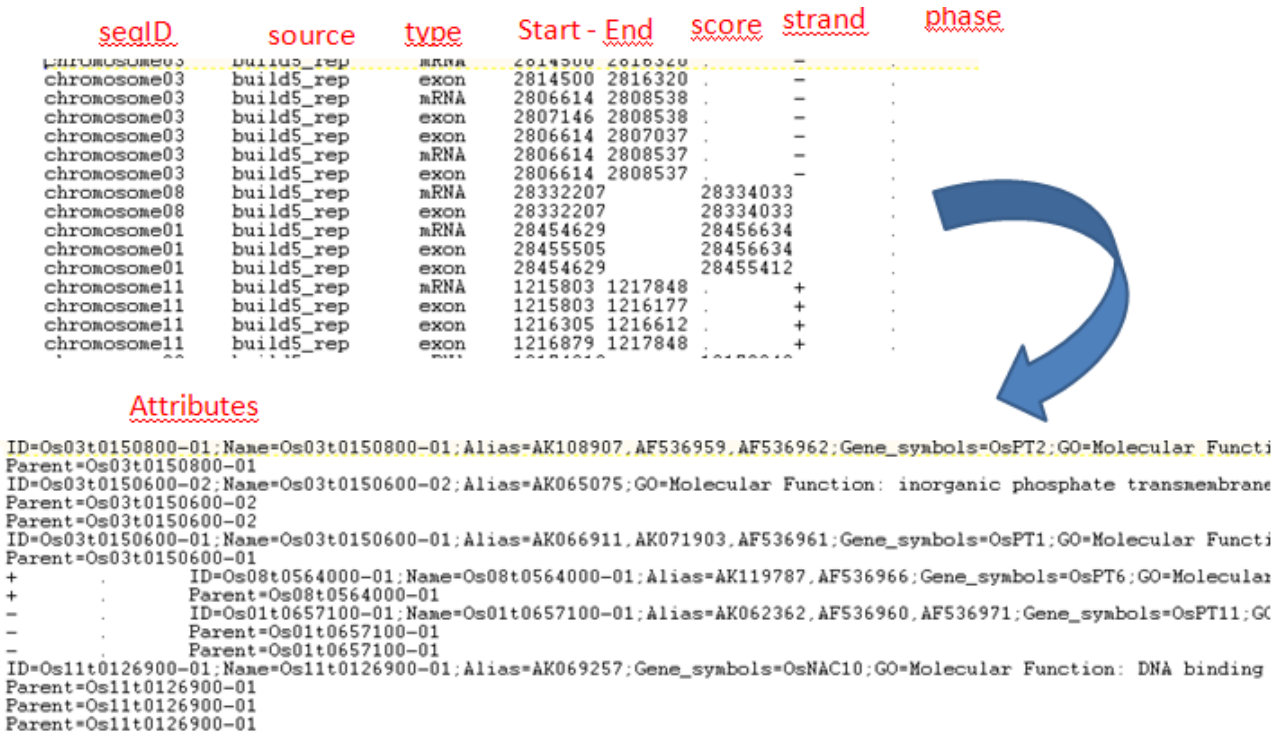


Figura 30 - Modelo padrão de um arquivo gff.

Contém os campos: seqID, source, type, start-end da sequência, score, strand (fita normal ou complementar), phase e attributes.

3ª etapa: Criação dos arquivos de índice, e se necessário, visualização das estatísticas do mapeamento com o software Samtools (Figura 31).

[6] Os nomes não foram traduzidos por serem nomes próprios dos campos do arquivo .SAM.

```
#! /bin/bash

export PATH=/usr/local/bin/:$PATH

#bams trimados e m = 2: /home/rafael/nobackup/home/Documentos/Oryza_sativa/mapeamento/osat_mapping/completed

#Rodar no seguinte caminho:/home/katia/nobackup/

bam="/home/rafael/nobackup/home/Documentos/Oryza_sativa/mapeamento/osat_mapping/completed/"

samtools index $bam/cr3bRun2_vs_chr06/accepted_hits.bam
samtools index $bam/cr3bRun2_vs_chr10/accepted_hits.bam
samtools index $bam/cr3bRun2_vs_chr05/accepted_hits.bam
```

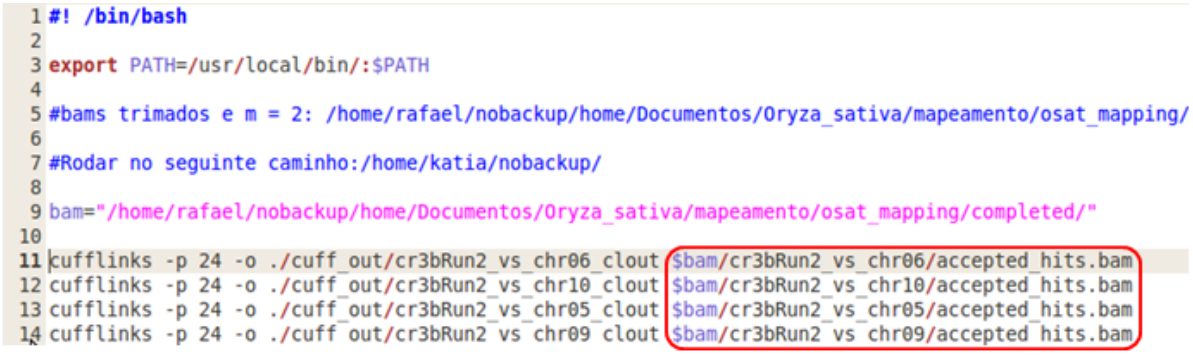
```
#! /bin/bash

export PATH=/usr/local/bin/:$PATH

#bams trimados e m = 2: /home/rafael/nobackup/home/Documentos/Oryza_sativa/mapeamento/osat_mapping/

#Rodar no seguinte caminho:/home/katia/nobackup/

bam="/home/rafael/nobackup/home/Documentos/Oryza_sativa/mapeamento/osat_mapping/completed/"

samtools idxstats $bam/cr3aRun1_vs_clor/accepted_hits.bam
samtools idxstats $bam/cr3aRun1_vs_mit/accepted_hits.bam
samtools idxstats $bam/cr3aRun2_vs_clor/accepted_hits.bam
samtools idxstats $bam/cr3aRun2_vs_mit/accepted_hits.bam
```

Figura 31 - Software Samtools.

Geração dos arquivos de índice e estatística do mapeamento com uso dos parâmetros samtools índex e samtools idxstats.

4ª etapa: Manipulação dos arquivos de anotação e de genoma de referência. É importante que o cabeçalho do arquivo FASTA seja idêntico à informação contida na primeira coluna dos arquivos GFF.

5ª etapa: Uso do software Cufflinks. A primeira etapa da execução do software Cufflinks cria os arquivos de transcritos para cada biblioteca, conforme descrito no item 1.6.3. A Figura 32 apresenta um modelo do script utilizado para tal finalidade. Em vermelho, são destacados os arquivos output contendo os alinhamentos no formato.BAM.



```
#! /bin/bash

export PATH=/usr/local/bin/:$PATH

#bams trimados e m = 2: /home/rafael/nobackup/home/Documentos/Oryza_sativa/mapeamento/osat_mapping/

#Rodar no seguinte caminho:/home/katia/nobackup/

bam="/home/rafael/nobackup/home/Documentos/Oryza_sativa/mapeamento/osat_mapping/completed/"

cufflinks -p 24 -o ./cuff_out/cr3bRun2_vs_chr06_clout $bam/cr3bRun2_vs_chr06/accepted_hits.bam
cufflinks -p 24 -o ./cuff_out/cr3bRun2_vs_chr10_clout $bam/cr3bRun2_vs_chr10/accepted_hits.bam
cufflinks -p 24 -o ./cuff_out/cr3bRun2_vs_chr05_clout $bam/cr3bRun2_vs_chr05/accepted_hits.bam
cufflinks -p 24 -o ./cuff_out/cr3bRun2_vs_chr09_clout $bam/cr3bRun2_vs_chr09/accepted_hits.bam
```

Figura 32 - Utilização da primeira etapa do software Cufflinks.

Aqui, são passados os arquivos de alinhamento .BAM para primeira etapa de montagem dos transcritos.

6ª etapa: Uso do pacote Cuffmerge do Cufflinks. Neste momento é feito a montagem dos transcritos, conforme especificado no arquivo assemblies.txt (destacado em vermelho na Figura 33). Nessa etapa é possível obter os dados de expressão.

```
#! /bin/bash

export PATH=/usr/local/bin/:$PATH

#anotação = /home/katia/nobackup/arroz_cuff_all/genoma/anotacao/
#fasta = /home/katia/nobackup/arroz_cuff_all/genoma/
#Deve ser executado no caminho: /home/katia/nobackup/cuff_out/

cuffmerge -p 24 -g osat_1a12.gff3 -s osat_1a12.fasta assemblies.txt
```

Figura 33 - Montagem dos transcritos com o pacote Cuffmerge.

Neste momento é feita segunda etapa de montagem dos transcritos, conforme especificado no arquivo assemblies.txt. O arquivo .fasta auxilia na montagem.

7ª etapa: Uso do pacote Cuffdiff do Cufflinks. O Cuffdiff é o último pacote do Cufflinks, onde são gerados os dados de expressão diferencial. São gerados dados de genes, isoformas, TSS, CDS, promoters, *splicing* e relCDS. A Figura 34 demonstra um exemplo do script gerado para essa etapa.

Figura 34 - Uso do Cuffdiff para criar os dados sobre expressão diferencial.

Aqui são passados os arquivos .fasta e .BAM novamente para ancoragem da expressão diferencial. Especifica-se ainda o nome das bibliotecas e o local do arquivo de junção dos transcritos (criado pelo Cuffmerge), por *default* chamado merged.gtf.

8ª etapa: Uso da biblioteca CummeRbund do R/Bioconductor para criação dos gráficos de interesse.

Posteriormente, foi criado um banco de dados na linguagem MySQL para manipulação dos dados. Este banco de dados possui, atualmente, 186 tabelas do projeto de arroz. Nem todas as tabelas da análise foram adicionadas à base de dados, somente aquelas necessárias para gerar os resultados sobre os transcritos novos e análise de cobertura dos dados. A Figura 35 apresenta um Diagrama de Entidade Relacionamento (DER) com as tabelas essenciais. O item 144_table_cov se refere a 144 tabelas com a mesma estrutura, mas com os dados de cobertura do arroz.

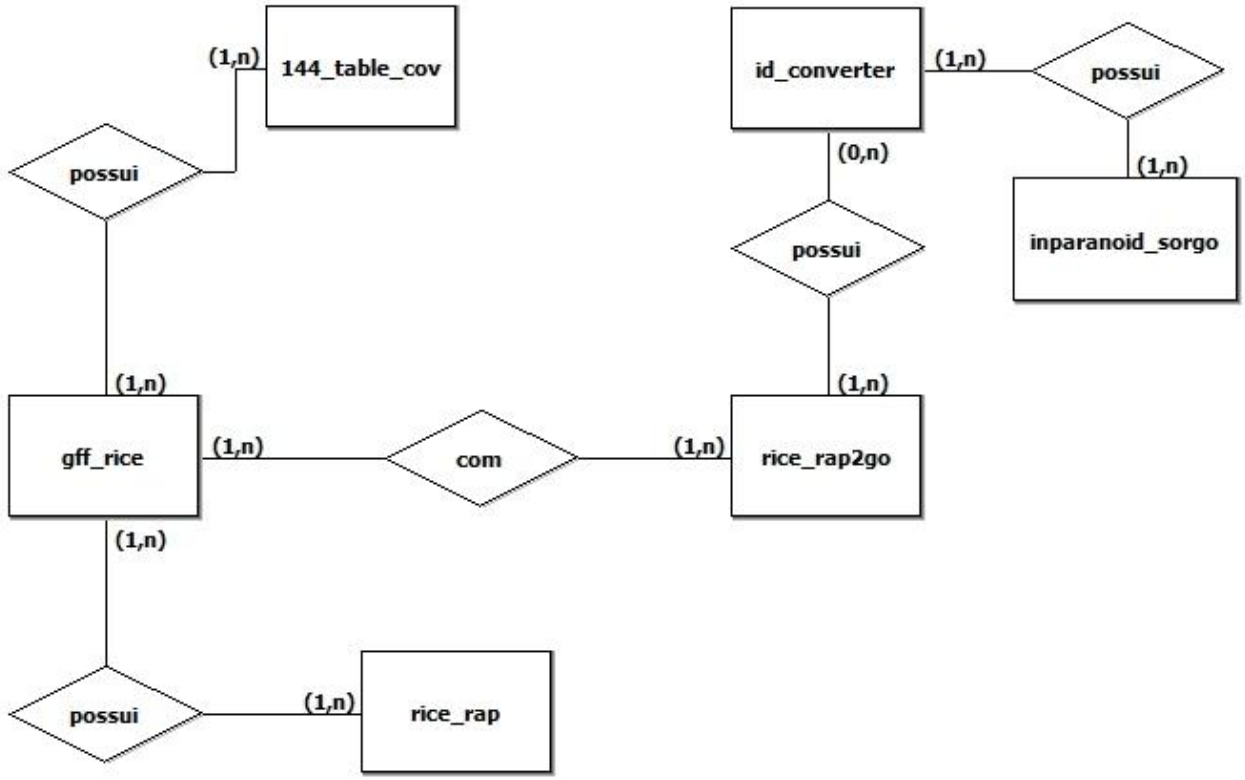


Figura 35 - Diagrama Entidade Relacionamento (DER) do banco de dados do transcriptoma de arroz.

A parte principal desta base de dados é relação das tabelas de identificação dos genes (criadas a partir dos arquivos de anotação dos cromossomos e das organelas) com as demais. Dessa maneira, foi possível trabalhar com diferentes ID's que foram atribuídos ao genoma do arroz por diferentes bases de dados disponíveis.

A parte principal deste banco de dados é a associação dos arquivos de anotação às várias tabelas do banco de dados. A tabela gff_rice contém o arquivo de anotação dos cromossomos do arroz, além desta existem também as tabelas de anotação das organelas: mitocôndria e cloroplasto. Foi criado um identificador (ID) para cada característica (*feature*) do gff, e as tabelas de cobertura (144 tabelas), rice_rap2go e rice_rap recebem esse ID como chave estrangeira, assim, é possível correlacionar os nomes de genes e saber a anotação rapidamente (conteúdo do campo CDS).

Este projeto utilizou dados públicos e alguns destes dados apresentaram algumas dificuldades: Existem pelo menos três tipos de identificadores para os genes de arroz e por isso, foi criado pelo RAP-DB uma ferramenta chamada

ID_CONVERTER[7]. Portanto, para que fosse possível trabalhar com os ID's do RAP-DB e os demais (como os ID's do site Inparanoid e MSU, por exemplo) foi necessário criar no banco de dados do arroz, uma tabela para conversão desses ID's denominada id_converter.

Contudo, a tabela inparanoid_sorgo possui pelo menos um identificador na tabela id_converter, que também está associada à tabela rice_rap2go. Assim, fica possível manipular os identificadores do Inparanoid, do RAP-DB e do GO facilmente. As demais tabelas, criadas pelo Cuffdiff não foram manipuladas neste banco de dados porque o R/Bioconductor cria um banco internamente em SQLite e a biblioteca CummeRbund acessa esses dados facilmente.

[7] Ferramenta ID_CONVERTER: < http://rapdb.dna.affrc.go.jp/tools/converter> Acesso em 06 de Nov. 2012.

4. RESULTADOS E DISCUSSÃO

Os dados foram analisados de duas maneiras: (1) Por cromossomo, com um índice para cada cromossomo do arroz, adicionando ainda as sequencias representativas das organelas: mitocôndria e cloroplasto. (2) Índice único contendo todos os cromossomos juntos. Foram analisados ainda, expressão para os dados de análise por cromossomo e os dados referentes ao TSS diferencialmente expressos tanto para índice único quanto para o múltiplo, além de uma análise geral dos transcritos novos.

Os itens que se segue descrevem os resultados obtidos.

4.1. ANÁLISE DO MAPEAMENTO

A Tabela 5 apresenta a quantidade de *reads* sequenciadas por biblioteca. Foi criado um gráfico para melhor visualização da distribuição das *reads* conforme Figura 36. Onde é possível verificar que a biblioteca cr7aRun1 possui o maior número de *reads*. E a biblioteca ir3bRun1 possui a menor quantidade de *reads*.

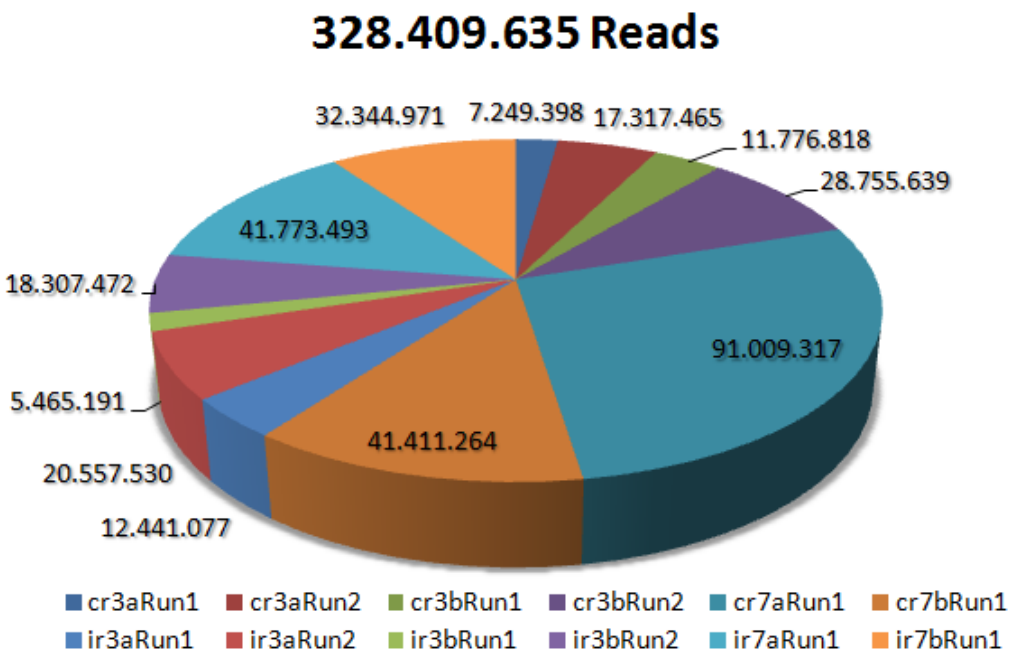


Figura 36 - Quantidade de *reads* por biblioteca.

Possível verificar a contribuição de cada biblioteca para as análises, onde, a biblioteca cr7aRun1 possui o maior número de *reads* e a biblioteca ir3bRun1 possui a menor quantidade de *reads.*

216.980.165 (75,2%) representam *reads* exônicas (Figura 37), e 71.493.386 (24,8%) mapearam em junções de éxons (Figura 38). Esses números totalizam 288.473.551 eventos de mapeamento, evidenciando que 85.702.567 dessas *reads* mapeavam contra mais de uma referência. Um total de 202.770.984 (61,7%) *reads* mapearam pelo menos uma vez contra o genoma do arroz.

Figura 37 - *Reads* exônicas do arroz. Representam um total de 216.980.165 (75,2%).

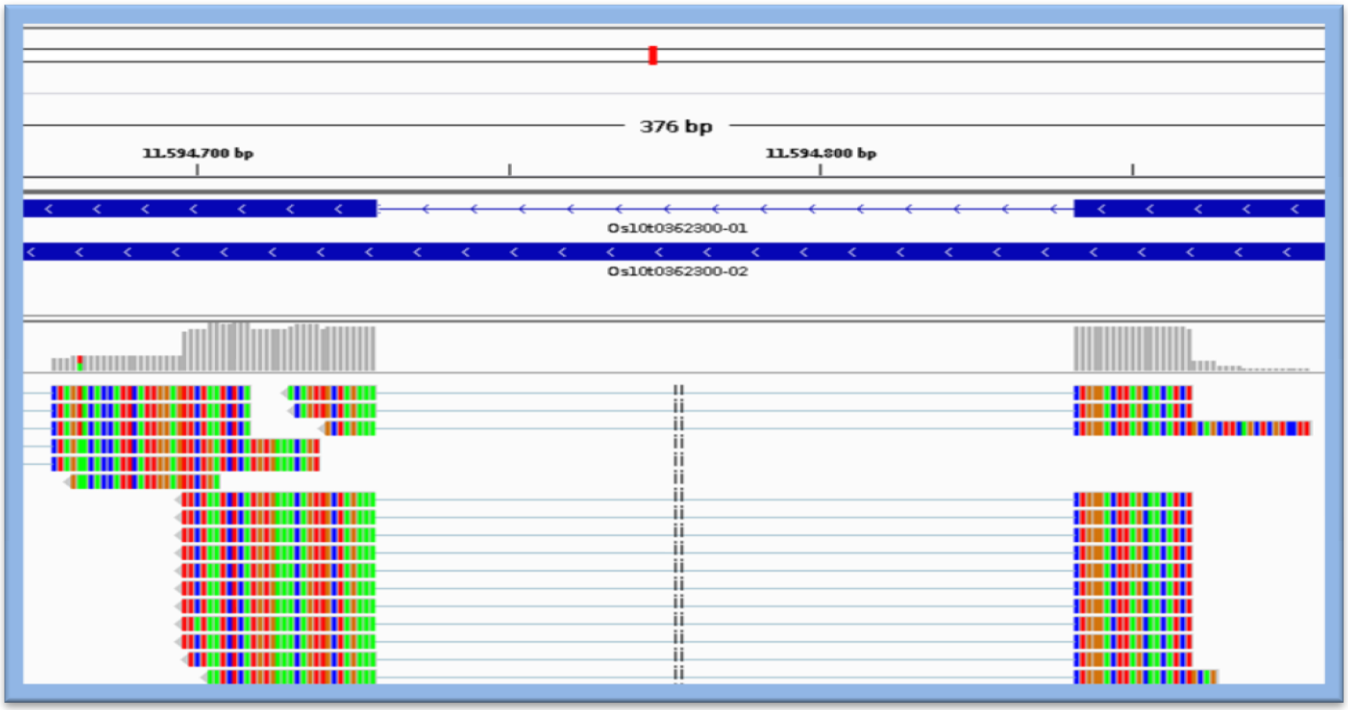


Figura 38 - Junções de éxons. Representados por 71.493.386 (24,8%) *reads*.

Um total de 6.942 (16,4%) e 4.915 (11,6%) genes de arroz obtiveram cobertura nas bibliotecas: CR3 e IR3; e 6.733 (15,9%) e 3.807 (9%) genes para as bibliotecas CR7 e IR7, respectivamente. A Figura 39 ilustra a cobertura das características gênicas, visando a identificação dos genes expressos.

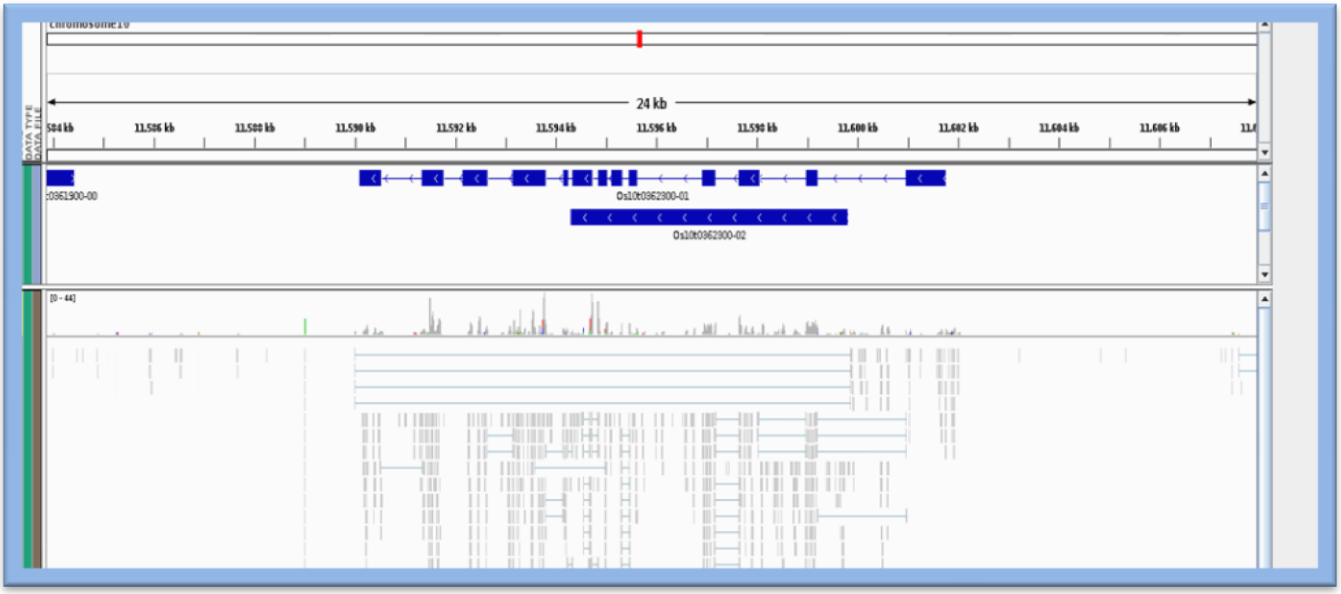


Figura 39 - Empilhamento de *reads* visualizados através do software IGV.

Ilustra a cobertura das características gênicas, visando a identificação dos genes expressos.

Para verificação de outras características do mapeamento do transcriptoma contra o genoma, foram realizados testes, e os principais são listados nas Tabelas 6 - 10.

A Tabela 6 demonstra um dos problemas encontrados. Utilizando o arquivo GFF sem alteração o Cufflinks não reconhece o nome dos genes, assim, todos seriam transcritos novos, o que não está correto. No entanto, a Tabela 7 mostra como esse problema foi resolvido. Foi necessário utilizar o software gffread do Cufflinks para entendermos como o Cufflinks considera as colunas do arquivo GFF e em seguida, foi necessário criar uma coluna unicamente com o nome dos genes.

Outro fator que também diferencia os resultados é o parâmetro de fita específica que deve ser utilizado com dados gerados pelo sequenciador SOLiD (Tabela 8). E por fim, foram verificados os dados de expressão das replicatas técnicas e biológicas do cromossomo 10 apenas para efeito de comparação (Tabelas 9 - 10).

Tabela 6 - Teste 1 com o cromossomo 10 de arroz, GFF sem alteração, valor de mismatch igual a 2 e versão 1.2.1 do Cufflinks.

Amostras	Transcritos totais	TSS	Genes D.E	TSS D.E	Genes RAP-DB	Genes anotados	Transcritos novos
2	1739	1963	95	92	2242	0	Todos*

Nota: * conforme explicado no texto.

Tabela 7 - Teste 2 com cromossomo 10 de arroz, GFF alterado manualmente, valor de mismatch igual a 2 e versão 1.2.1 do Cufflinks.

Amostras	Transcritos totais	TSS	Genes D.E	TSS D.E	Genes RAP-DB	Genes anotados	Transcritos novos
2	1739	1963	95	92	2242	838	901

Tabela 8 - Teste 3 com cromossomo 10 de arroz, GFF alterado, versão 1.2.1 do Cufflinks e parâmetro de fita específica do TopHat.

Amostras	Transcritos totais	TSS	Genes D.E	TSS D.E	Genes RAP-DB	Genes anotados	Transcritos novos
2	1685	1874	209	203	2242	679	1006

Tabela 9 - Teste 4 para verificação dos resultados de expressão das replicatas técnicas e biológicas. Diferenciando 3 e 7 dias, controle e inoculado do cromossomo 10 de arroz.

Amostras	Condições	Replicatas	Transcritos totais	TSS	Genes D.E	TSS D.E	Genes RAP-DB	Genes anotados	Transcritos novos
2	3dias_controle	Run1 e 2	949	1021	260	254	2242	386	563
2	3dias_inoculado	Run1 e 2	861	897	259	256	2242	274	587
2	7dias_controle	a e b	1906	1938	86	0	2242	418	1488
2	7dias_inoculado	a e b	1048	1058	73	0	2242	196	852

Tabela 10 - Teste 5 para verificação dos dados de replicatas técnicas. Cromossomo 10 de arroz, bibliotecas controle e inoculado de 3 dias das Rodadas 1 e 2 do SOLiD.

Amostras	Condições	Transcritos totais	TSS	Genes D.E	TSS D.E	Genes RAP-DB	Genes anotados	Transcritos novos
2	cr3a1 vs. cr3a2	590	611	154	158	2242	167	423
2	cr3b1 vs. cr3b2	872	909	238	241	2242	325	547
2	ir3a1 vs. ir3a2	537	555	201	205	2242	168	369
2	ir3b1 vs. ir3b2	753	764	248	248	2242	189	564

Após a realização de todos os testes foi possível chegar a um consenso de utilização dos parâmetros dos softwares Bowtie, TopHat e Cufflinks para o mapeamento do arroz. Foi resolvido inclusive, a utilização da versão 1.2.1 do

Cufflinks por ser a mesma publicada no artigo da revista *Nature* (TRAPNELL, ROBERTS, *et al.*, 2012).

Após a realização de todos os procedimentos apresentados é possível visualizar o mapeamento com uso do Software IGV do *Broad Institute* (ROBINSON, THORVALDSDÓTTIR, *et al.*, 2011). A Figura 40 é o mapeamento feito sem o arquivo de anotação. A Figura 41 demonstra a necessidade de utilização do parâmetro --library-type, porque sem ele, *reads* que mapeiam em ambas as fitas podem estar sujeitas a montagem como um único transcrito (não mostrado neste exemplo). As Figuras 42 e 43 demostram o mapeamento das *reads* em fita específica e com o arquivo de anotação para ancoragem na montagem dos transcritos.

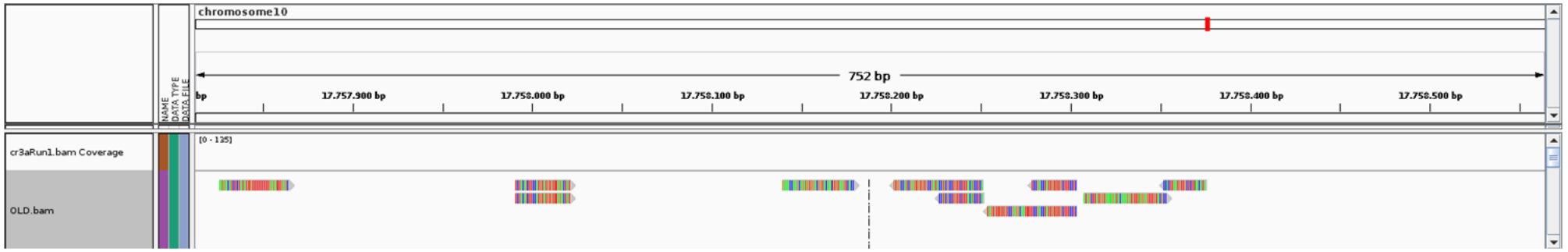


Figura 40 - Mapeamento feito sem o arquivo de anotação.

Neste caso é possível observar a orientação das *reads* nos dois sentidos. Sem a ancoragem do arquivo de anotação, o Cufflinks deixa de montar transcritos que não possuam um empilhamento de *reads* considerável.

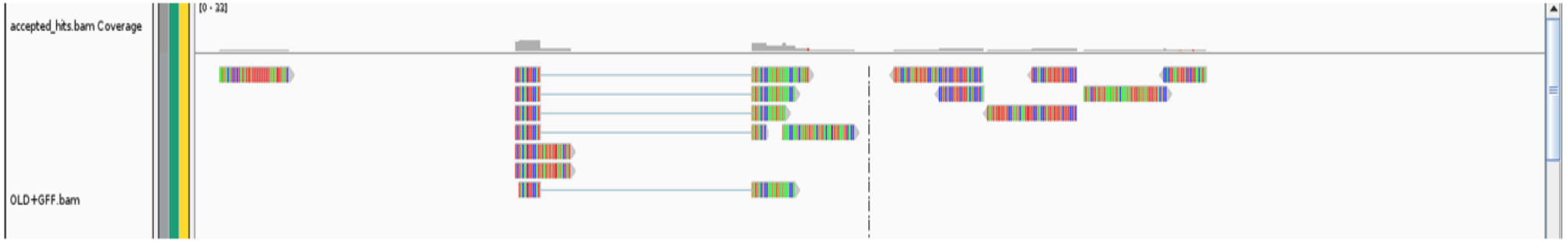


Figura 41 - Mapeamento nas duas fitas do DNA.

Demonstra a necessidade de utilização do parâmetro --library-type do Cufflinks, porque sem ele, *reads* que mapeiam em ambas as fitas podem estar sujeitas a montagem como um único transcrito (não mostrado neste exemplo).

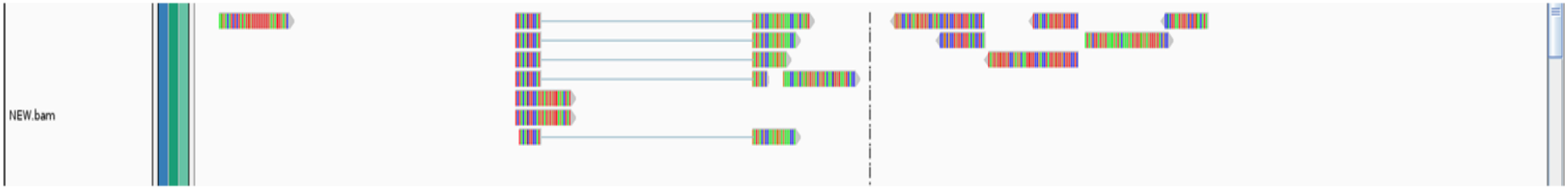


Figura 42 - Mapeamento em fita específica.

Este exemplo foi feito com um mapeamento em que foi utilizado o parâmetro –library-type. Assim, somente *reads* na mesma orientação estão sujeitas à montagem como um único transcrito.

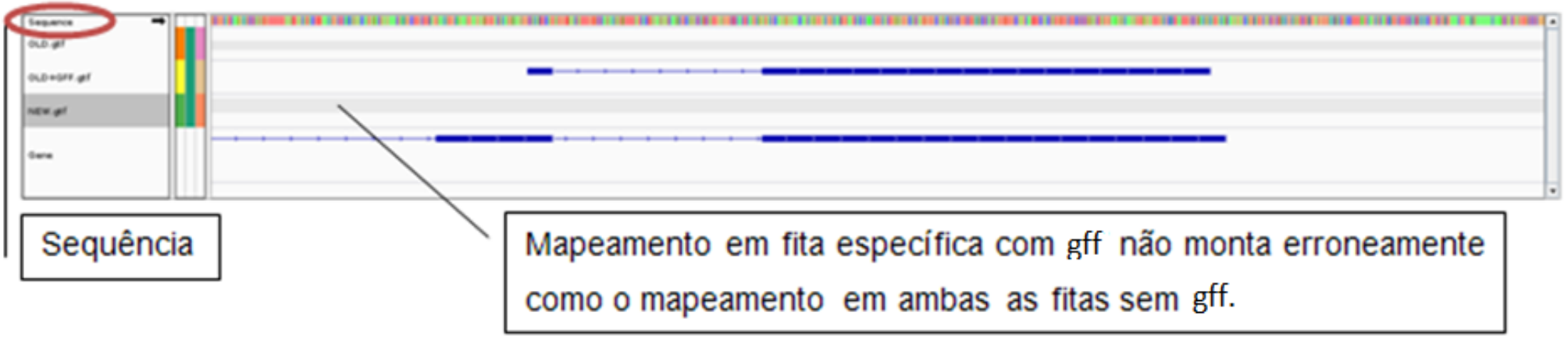


Figura 43 - Mapeamento em fita específica com arquivo de anotação gff.

A primeira barra horizontal azul indica a montagem antiga, sem arquivo gff e parâmetro de fita específica, onde o transcrito era montado erroneamente. Na linha abaixo (indicado pela caixa de texto), com uso do gff e do parâmetro --library-type, essa montagem não ocorre, demonstrando a necessidade de utilização do parâmetro de fita específica para *reads* geradas no Solid, além da utilização do arquivo de anotação para suporte.

Também é possível visualizar a sobreposições das *reads* para montagem dos transcritos em cada tratamento (Figura 44). Considerando a etapa de busca por expressão diferencial, é possível visualizar um transcrito fictício, gerado a partir de *reads* criadas para diferentes bibliotecas, que é utilizado para representar um *locus* gênico durante a análise de expressão diferencial (Figura 45). Esta Figura demonstra uma importante característica da montagem dos transcritos para estas análises, que é a utilização do arquivo de anotação GFF para dar suporte a determinação dos transcritos, um processo chamado: RABT (*Reference Annotation Based Transcript*). Ao montar o transcrito, o Cuffmerge une alguns éxons que não são necessariamente sobrepostos, uma vez que o arquivo GFF funciona como uma "ponte" adicional para definição das unidades de transcrição.

Figura 44 – Transcritos por biblioteca juntamente com arquivo de anotação.

Também é possível visualizar a sobreposições das *reads* para montagem dos transcritos em cada tratamento.

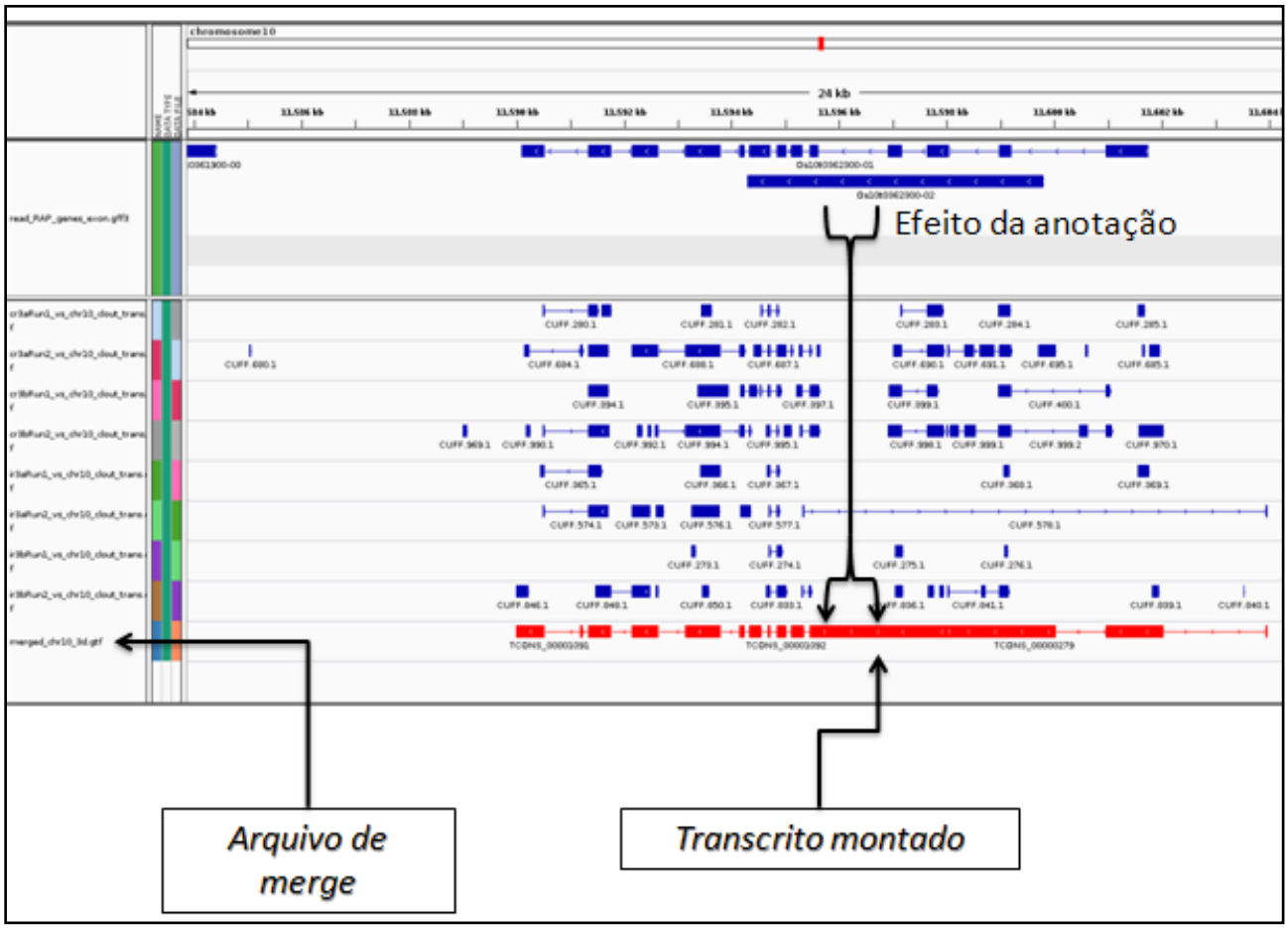


Figura 45 - Efeito do arquivo de anotação para montagem dos transcritos.

Considerando a etapa de busca por expressão diferencial, é possível visualizar um transcrito fictício, gerado a partir de *reads* criadas para diferentes bibliotecas, que é utilizado para representar um *locus* gênico durante a análise de expressão diferencial. Ao montar o transcrito, o Cuffmerge une alguns éxons que não são necessariamente sobrepostos, uma vez que o arquivo GFF funciona como uma "ponte" adicional para definição das unidades de transcrição.

Para análise de expressão diferencial os experimentos utilizados foram as amostras controle (CR) contra as amostras inoculadas (IR) pela *H. seropedicae*, todos agrupados no mesmo arquivo de montagem denominado "assemblies.txt": ou seja, IR + CR. No entanto, para testar a expressão diferencial, também foi realizada uma análise utilizando as bibliotecas em separado: apenas CR e apenas IR, conforme demonstrado na Figura 46. O arquivo "merged.gtf" gerado pelo Cuffmerge contêm o identificador das unidades de transcrição definidas TCONS_ID, que correspondem aos identificadores CUFF_ID presentes nos arquivos finais de montagem dos transcritos: "transcripts.gtf" (Figura 47). A Figura 48 demonstra como é representado o mapeamento quando é encontrado um transcrito novo. Estes transcritos não possuem uma região correspondente no arquivo de anotação, porém, o empilhamento de *reads* evidencia a existência do transcrito mesmo sem o suporte da anotação genômica presente no arquivo GFF.

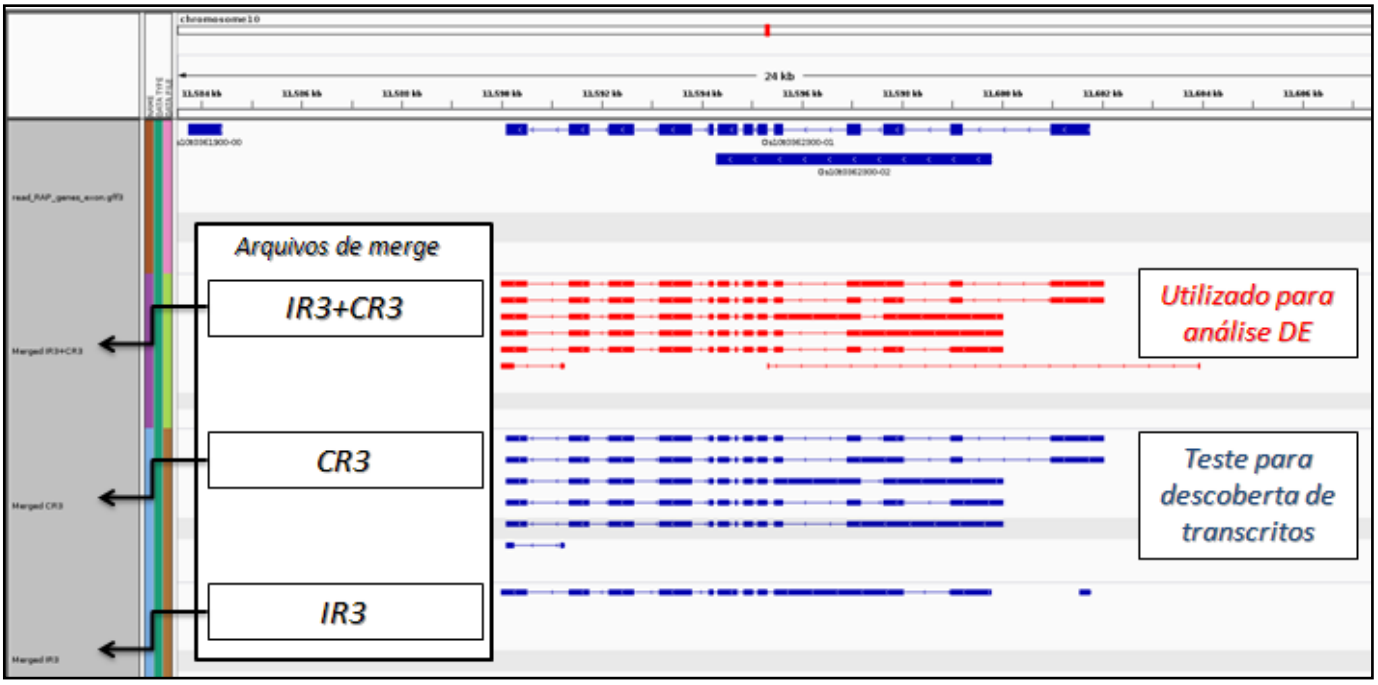


Figura 46 – Representação gráfica da montagem dos transcritos.

Montagem conjunta usando transcritos de ambos os tratamentos (em vermelho) utilizada para análise de expressão diferencial e montagens em separado (em azul) usada para análise de descoberta de transcritos.

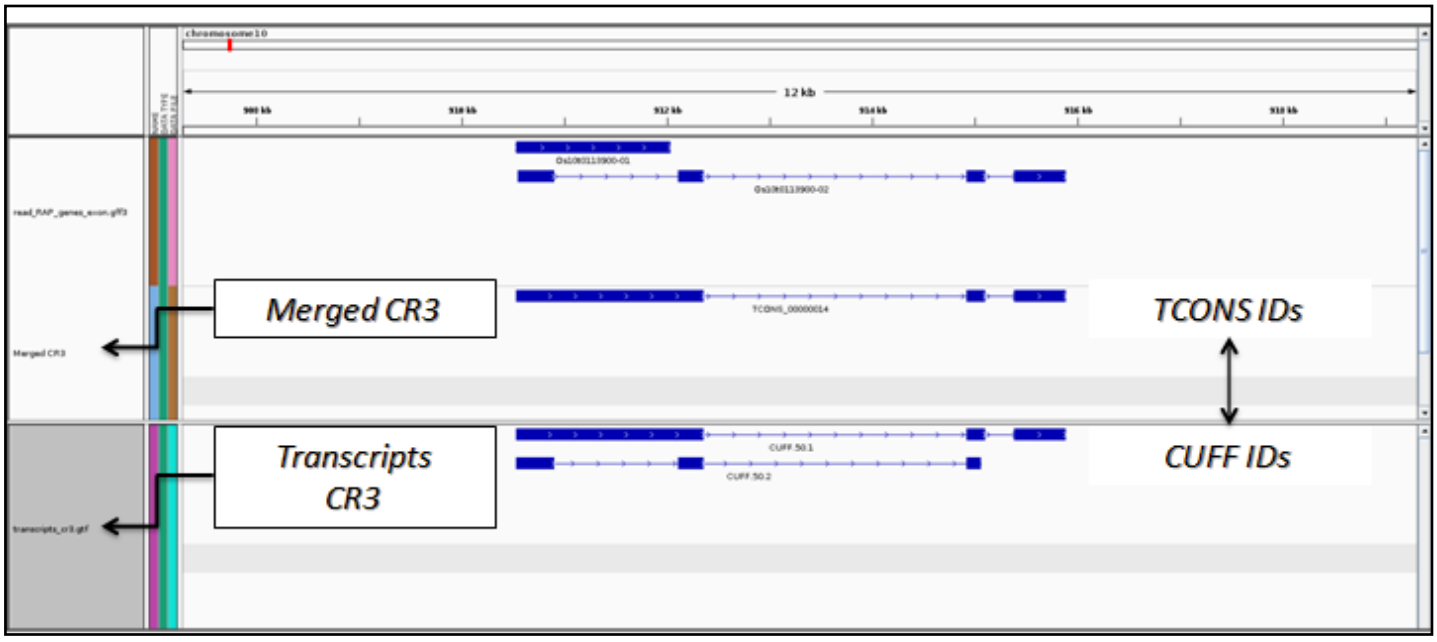


Figura 47 - Identificadores dos arquivos de merged e transcripts.

O arquivo “merged.gtf” gerado pelo Cuffmerge contêm o identificador das unidades de transcrição definidas TCONS_ID, que correspondem aos identificadores CUFF_ID presentes nos arquivos finais de montagem dos transcritos: “transcripts.gtf”. Na figura apresenta-se TCONS ID’s e CUFF ID’s respectivamente.

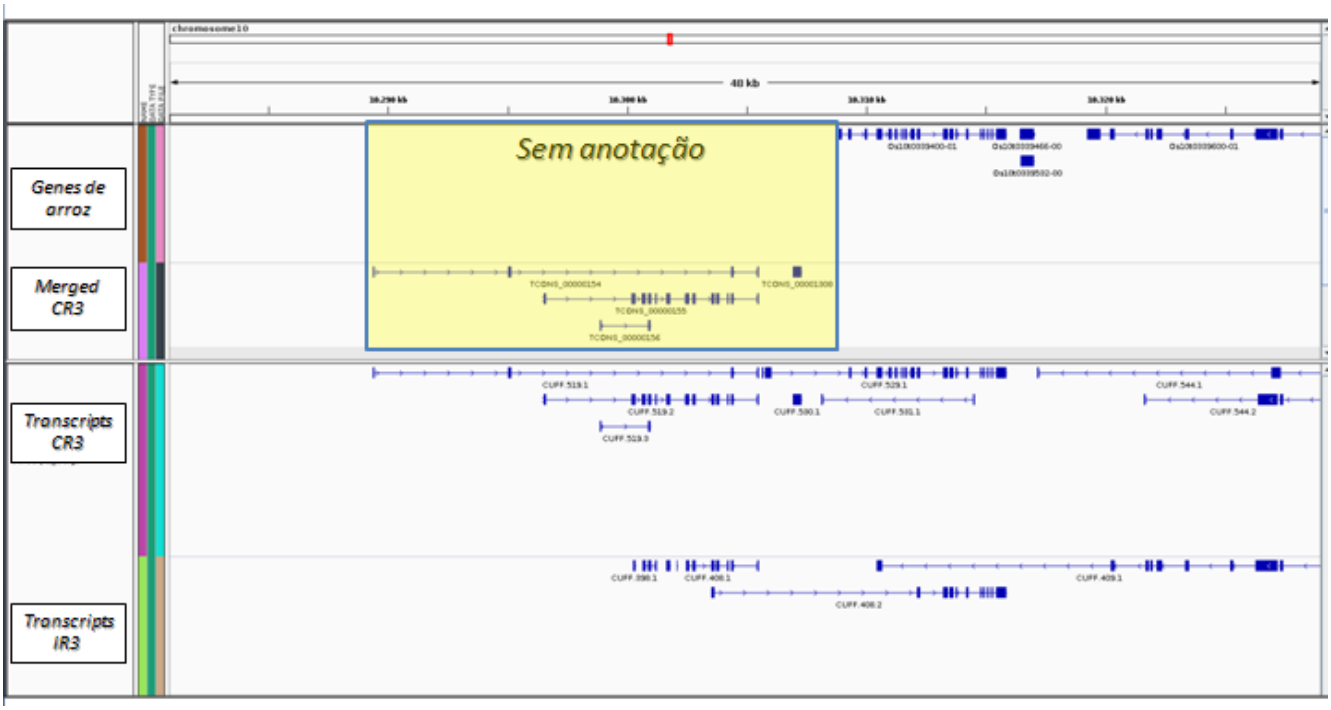


Figura 48 - Visualização de um transcrito novo.

Nota-se que no genoma de referência não existe uma região que contenha um gene anotado, porém, o empilhamento de *reads* evidencia a existência do transcrito mesmo sem o suporte da anotação genômica presente no arquivo GFF.

Foram obtidos alinhamentos para um total de 202.770.984 *reads* (média total de 61,74%) obtidas para o transcriptoma de arroz inoculado com a bactéria *H. seropedicae* contra o genoma de referência do arroz depositado na base de dados RAP-DB. A Figura 49 mostra um gráfico com os percentuais de *reads* mapeadas obtidos por biblioteca.

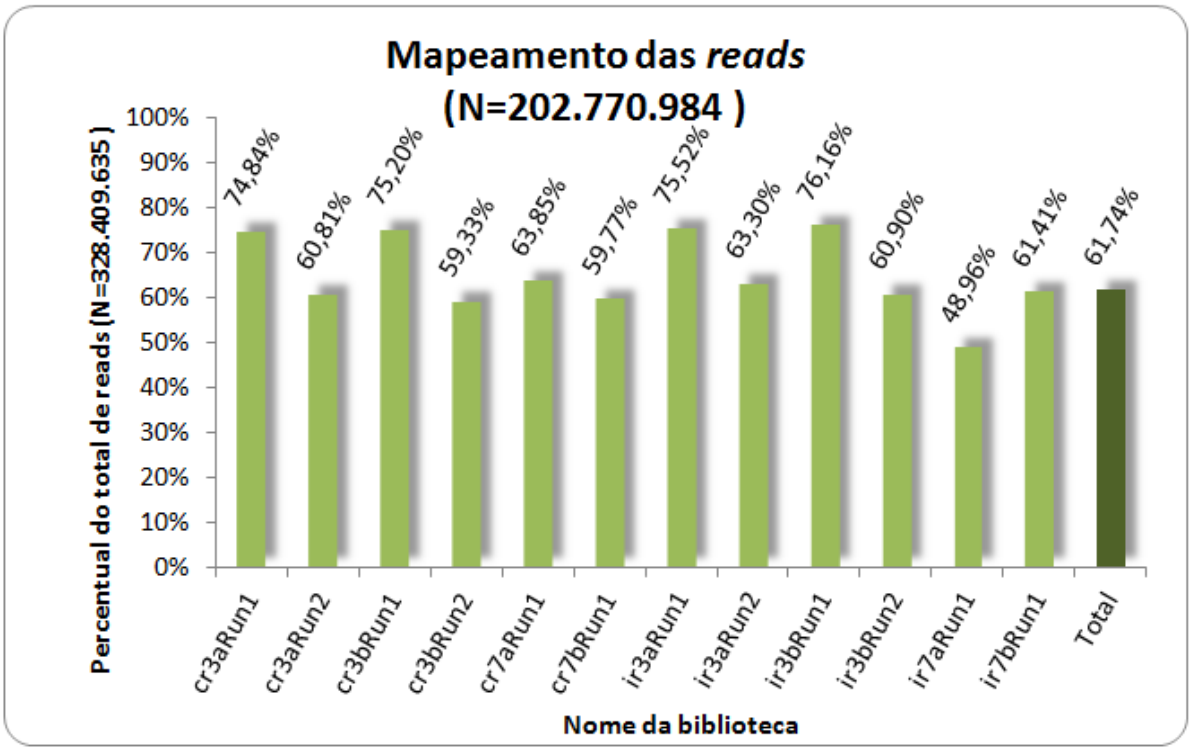


Figura 49 - Percentual de mapeamento por biblioteca.

Foram obtidos alinhamentos para um total de 202.770.984 *reads* (61,74%) obtidas para o transcriptoma de arroz inoculado com a bactéria *H. seropedicae* contra o genoma de referência do arroz depositado na base de dados RAP-DB.

As Figuras 50 e 51 apresentam gráficos com uma análise de recuperação de componentes representada por uma curva ROC. Esta curva representa o número total de *reads* mapeáveis por biblioteca recuperada em cada uma das condições do eixo X. Este eixo representa o número de hits das reads SOLiD contra cada um dos cromossomos individuais (nota-se que é válido apenas um *hit* por cromossomo), cada abscissa ($0 \geq x \leq 12$) possui como ordenada (y) o percentual recuperado do total de *reads* mapeáveis considerando x hits contra os cromossomos individuais. São apresentadas uma curva para cada biblioteca, onde a linha preta pontilhada representa a média entre as bibliotecas. Esta linha indica que 75,31% das *reads* mapeáveis obtidas para o tratamento de 3 dias são alinhadas contra apenas 1 cromossomo do genoma, enquanto esse valor sobe para 80,75% das *reads* mapeáveis obtidas para 7 dias de tratamento.

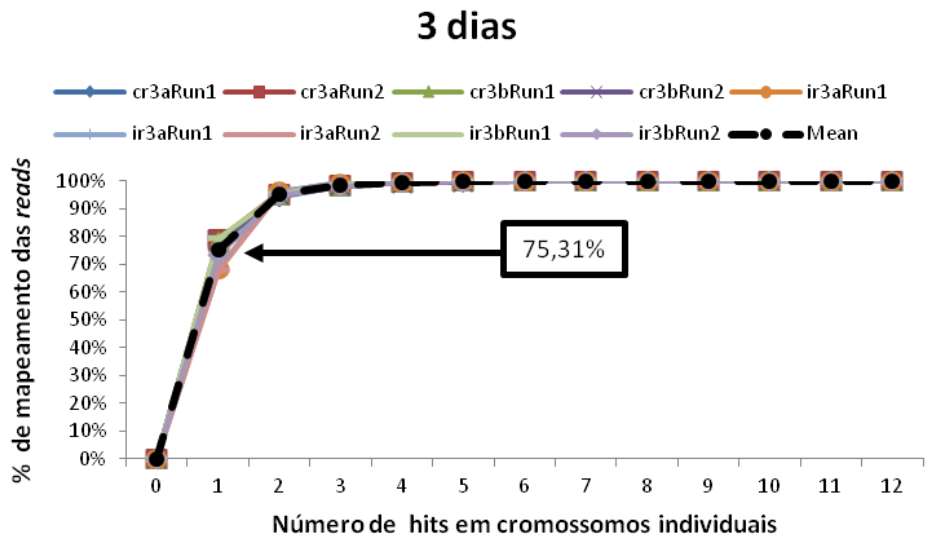


Figura 50 - Percentual de *reads* mapeáveis recuperadas nas bibliotecas de 3 dias considerando o número hits contra cromossomos individuais.

A linha preta pontilhada indica a média entre as bibliotecas, evidenciando que 75,31% das *reads* mapeáveis para o tratamento de 3 dias são alinhadas contra apenas 1 cromossomo do genoma.

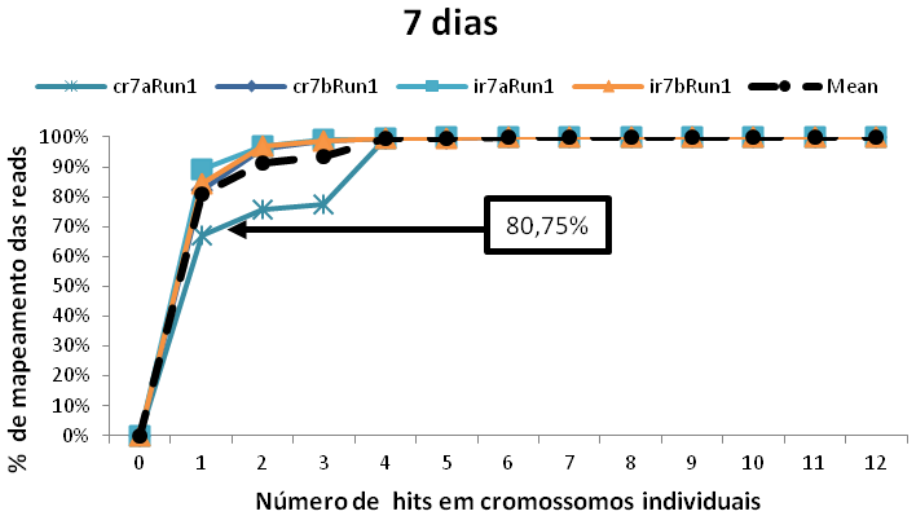


Figura 51 - Percentual de reads mapeáveis recuperadas nas bibliotecas de 7 dias considerando o número hits contra cromossomos individuais.

A linha preta pontilhada indica a média entre as bibliotecas, evidenciando que 80,75% das *reads* mapeáveis para o tratamento de 7 dias são alinhadas contra apenas 1 cromossomo do genoma.

Avaliamos também a obtenção de mapeamentos válidos (apenas 1 hit contra cromossomos individuais) em cada uma das bibliotecas considerando a contribuição desta para o número total de *reads* analisadas. Nota-se na Figura 52 que a biblioteca com maior contribuição percentual do número de *reads* (cr7aRun1) não foi necessariamente aquela que obteve o maior percentual de *reads* mapeadas, por isso a importância de se considerar as bibliotecas que contribuem com uma quantidade relativamente pequena de *reads*, pois estas podem contribuir significativamente para as inferências biológicas da análise, considerando o conjunto total de *reads* obtidas.

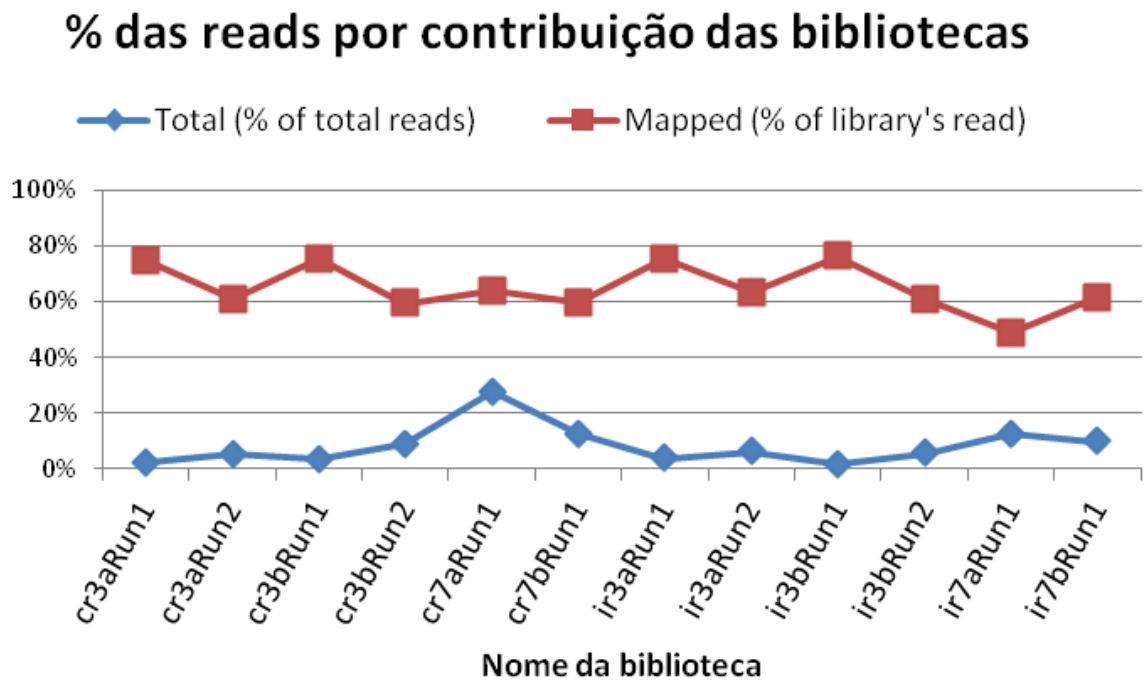


Figura 52 - Percentual de mapeamento por contribuição das bibliotecas.

Nota-se que a biblioteca com maior contribuição percentual do número de *reads* (cr7aRun1) não foi necessariamente aquela que obteve o maior percentual de *reads* mapeadas.

4.2. ANÁLISE DOS TRANSCRITOS MAPEADOS

A partir da obtenção dos arquivos de merged.gtf e transcripts.gtf resultantes da análise realizada com o programa Cuffmerge, foi possível realizar uma investigação sobre quais genes encontravam-se expressos para as bibliotecas de 3 e 7 dias, conforme apresentado nos gráficos das Figuras 53 - 56. Os resultados correspondem a transcritos mapeados, anotados, não-anotados, e que sofrem *splicing* alternativo conforme a distribuição apresentada na Tabela 11. Nota-se nos gráficos que os valores referentes aos genes localizados na mitocôndria e no cloroplasto são comparativamente elevados em relação aos valores obtidos para aqueles mapeados em outros cromossomos. Entretanto, ressalva-se que a quantidade de genes anotados nas sequências destas organelas é bem menor do que a dos outros cromossomos.

Tabela 11 – Distribuição dos transcritos identificados por amostra

Transcritos	**Amostras**			
	CR3	**IR3**	**CR7**	**IR7**
Mapeados	17.817	14.661	29.357	16.826
Anotados	9.523	6.193	8.418	4.452
Não-anotados	8.294	8.468	20.939	12.374
Splicing alternativo	9.382	6.035	8.206	4.315

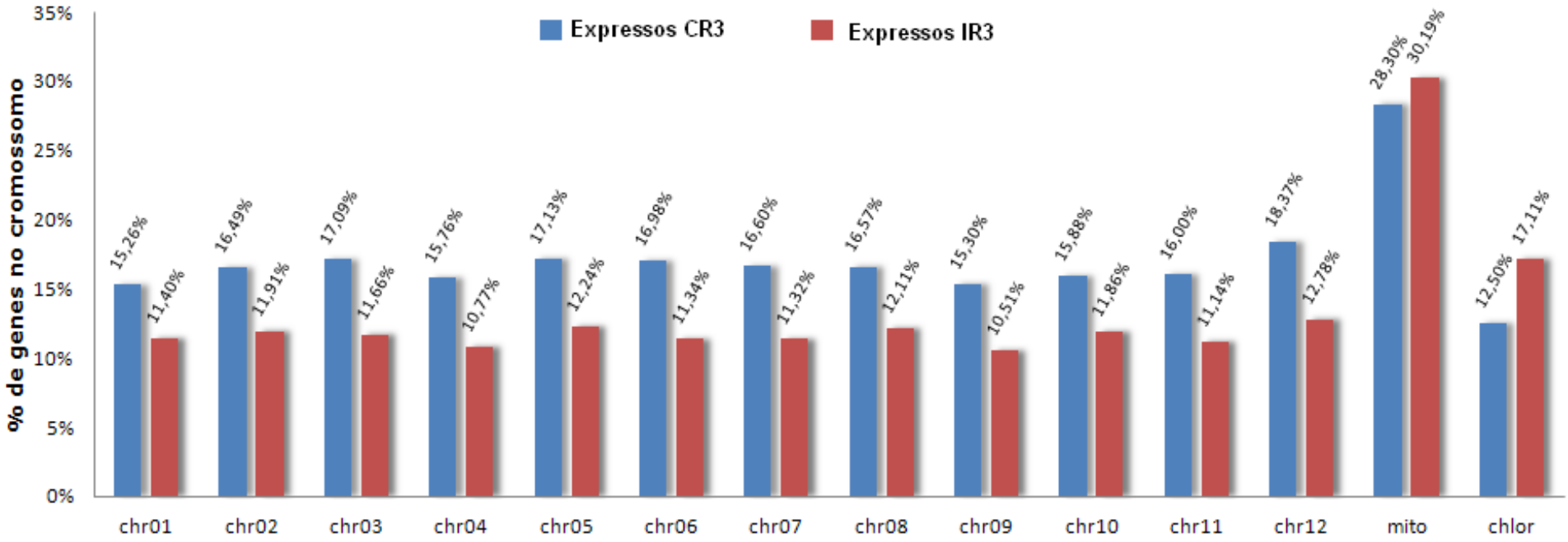


Figura 53 - Percentual de genes expressos em cada cromossomo nas bibliotecas CR3 e IR3.

Nota-se que os valores referentes aos genes localizados nas organelas (mitocôndria e cloroplasto) são elevados em comparação aos valores obtidos por cromossomo. Entretanto, ressalva-se que a quantidade de genes anotados nas sequências destas organelas é bem menor do que a dos outros cromossomos.

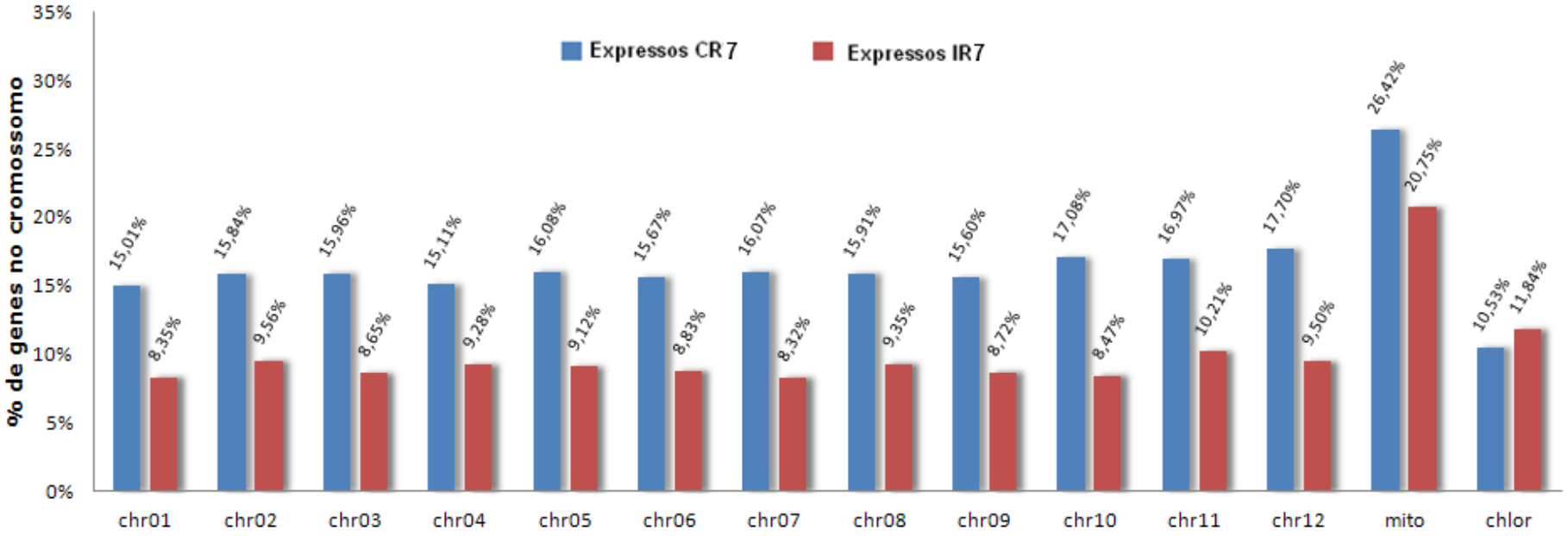


Figura 54 - Percentual de genes expressos em cada cromossomo nas bibliotecas CR7 e IR7.

Nota-se que os valores referentes aos genes localizados nas organelas (mitocôndria e cloroplasto) são elevados em comparação aos valores obtidos por cromossomo. Entretanto, ressalva-se que a quantidade de genes anotados nas sequências destas organelas é bem menor do que a dos outros cromossomos.

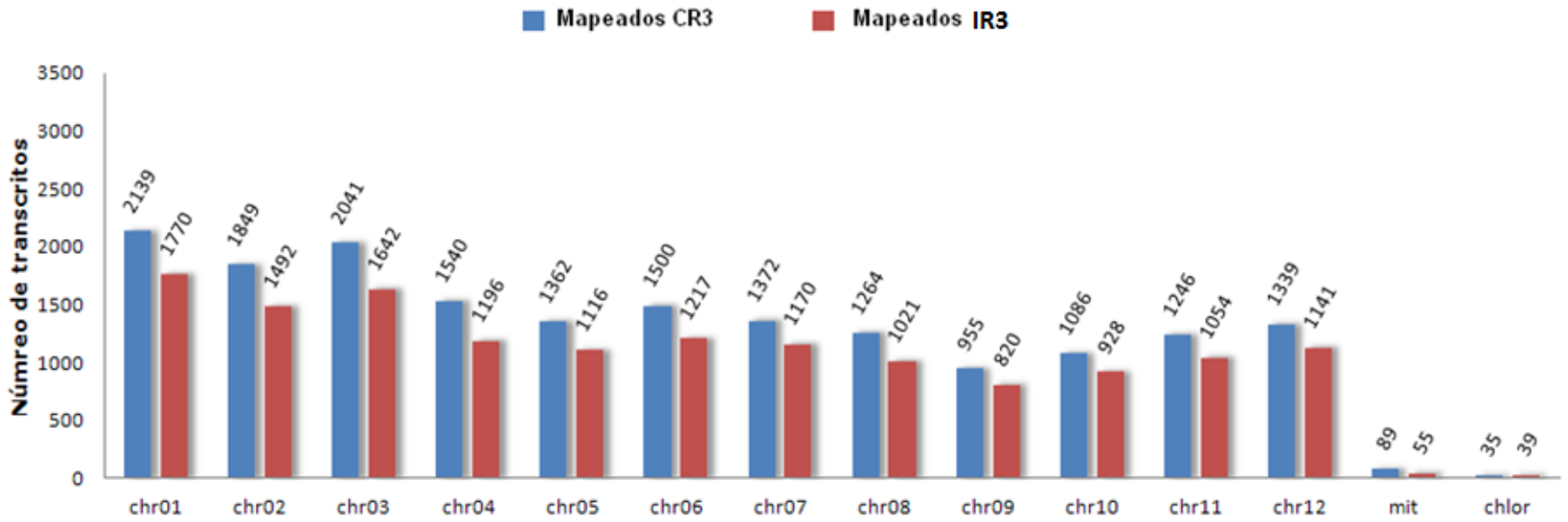


Figura 55 - Total de transcritos mapeados nas bibliotecas CR3 e IR3 por cromossomo.

Dados obtidos a partir dos arquivos de merged.gtf e transcripts.gtf resultantes da análise realizada com o programa Cuffmerge. Possível verificar que os cromossomos 1 e 3 apresentaram o maior número de transcritos mapeados tanto nas bibliotecas CR3 como IR3.

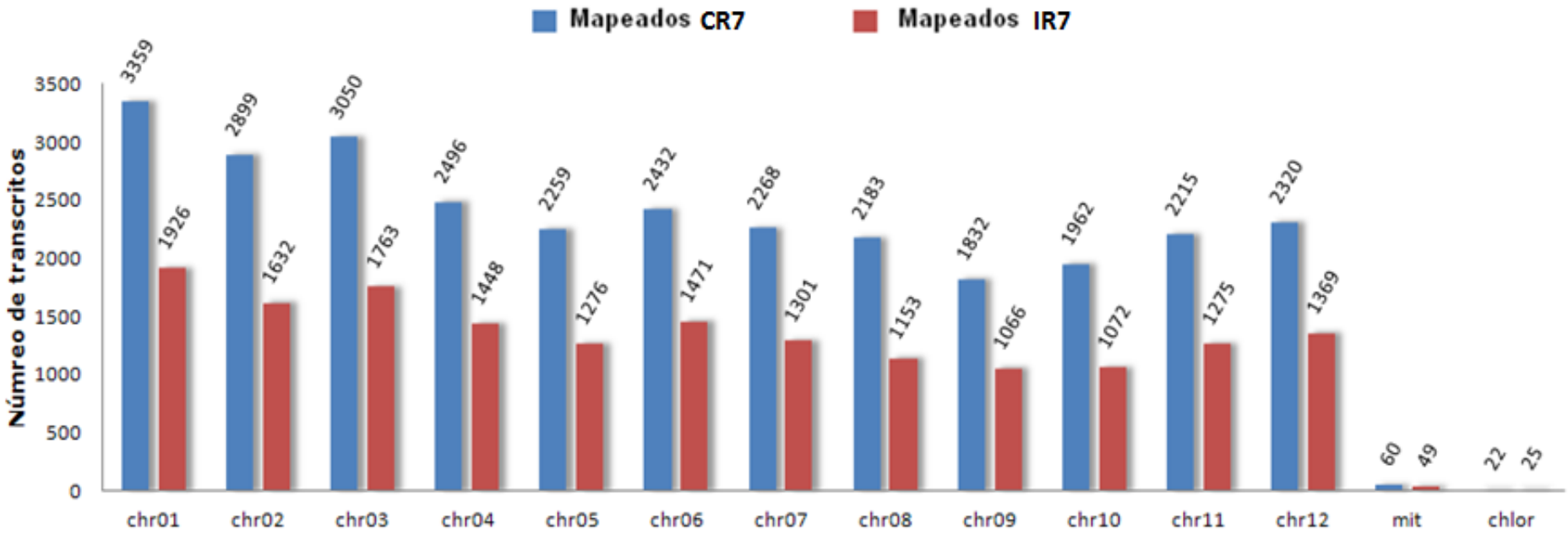
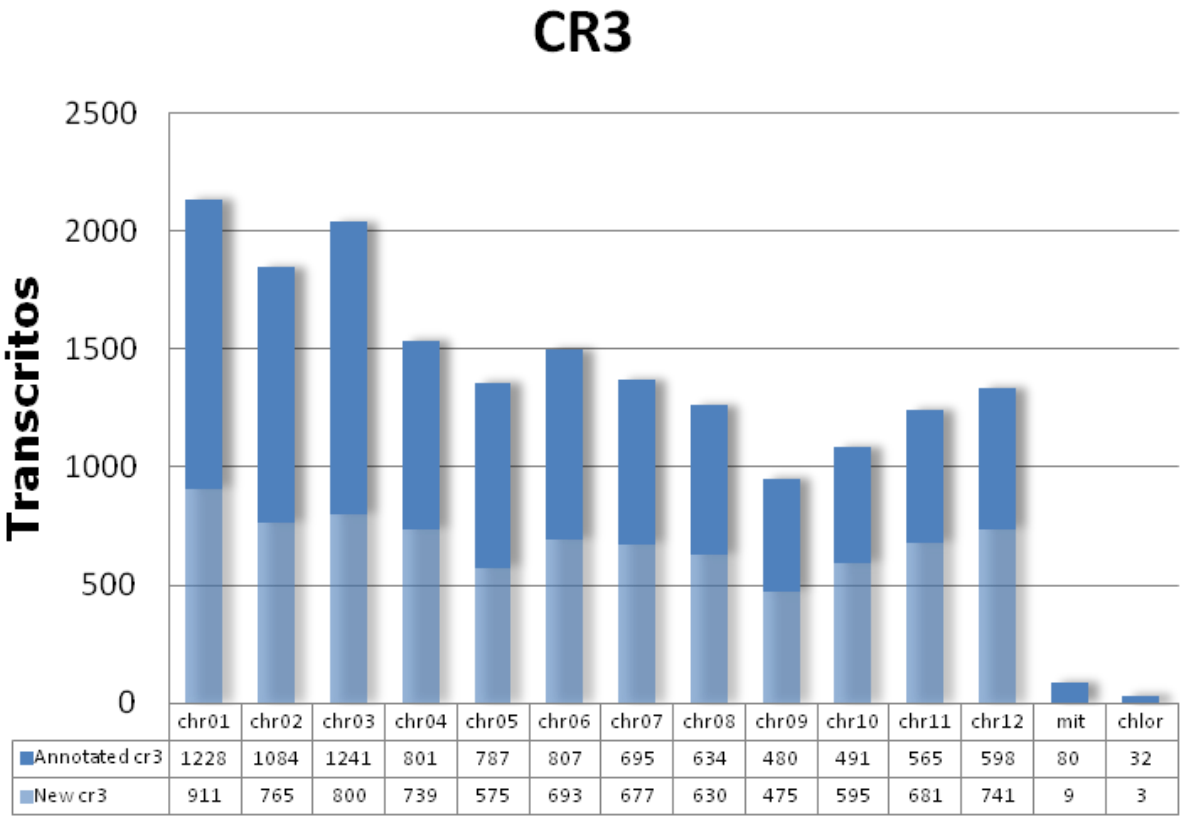


Figura 56 - Total de transcritos mapeados nas bibliotecas CR7 e IR7.

Dados obtidos a partir dos arquivos de merged.gtf e transcripts.gtf resultantes da análise realizada com o programa Cuffmerge. Após 7 dias de inoculação as diferenças em relação aos transcritos mapeados aumentam quando comparadas as bibliotecas controle e inoculado.

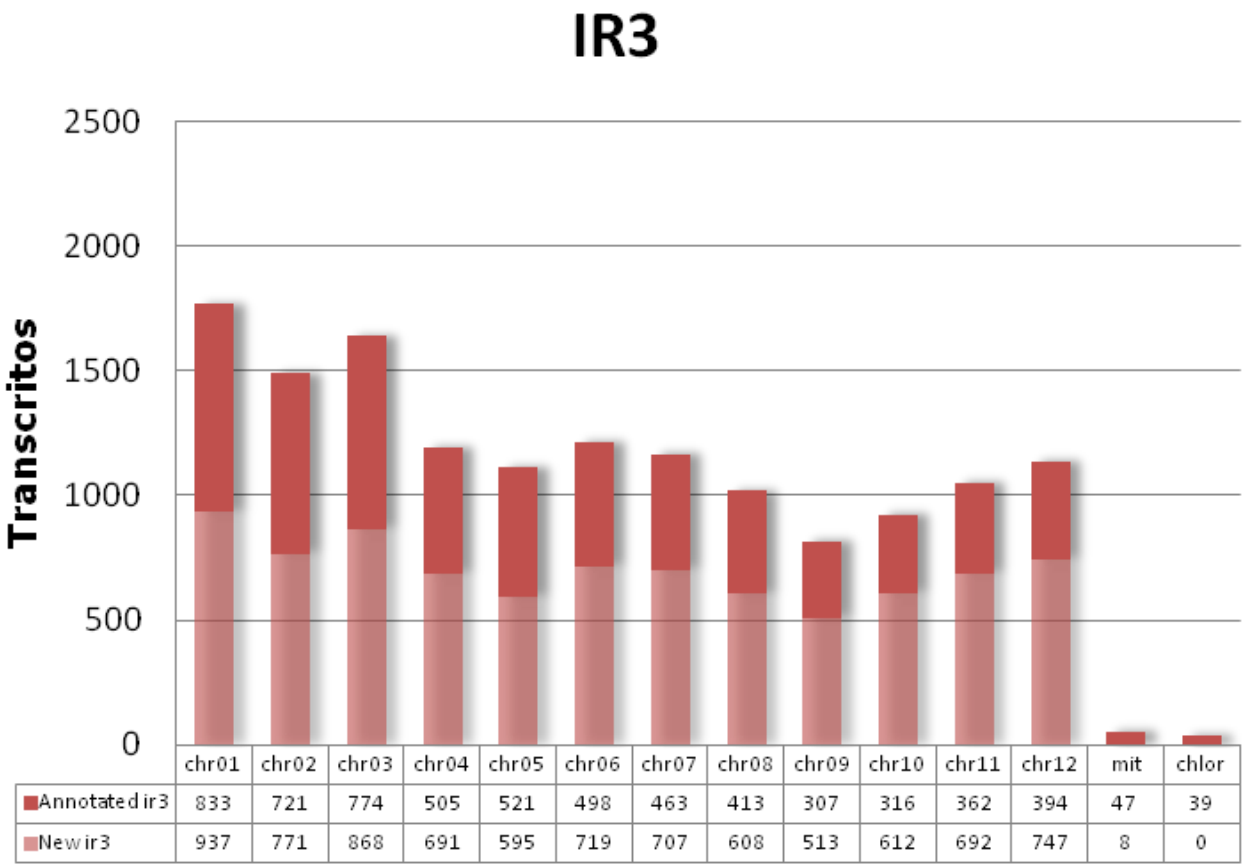
Foi possível ainda, quantificar o número de transcritos novos e o número de genes anotados expressos nas bibliotecas de controle e inoculado para 3 e 7 dias (Figuras 57 - 60). Os gráficos controles estão em azul e os dados referentes às bibliotecas inoculadas em vermelho, sendo que a altura das barras claras representa os transcritos novos enquanto a altura das barras escuras representa os genes anotados (vide tabela de dados abaixo dos respectivos gráficos).



	chr01	chr02	chr03	chr04	chr05	chr06	chr07	chr08	chr09	chr10	chr11	chr12	mit	chlor
Annotated cr3	1228	1084	1241	801	787	807	695	634	480	491	565	598	80	32
New cr3	911	765	800	739	575	693	677	630	475	595	681	741	9	3

Figura 57 - Relação entre transcritos anotados e transcritos novos para as bibliotecas controle de 3 dias, por cromossomo.

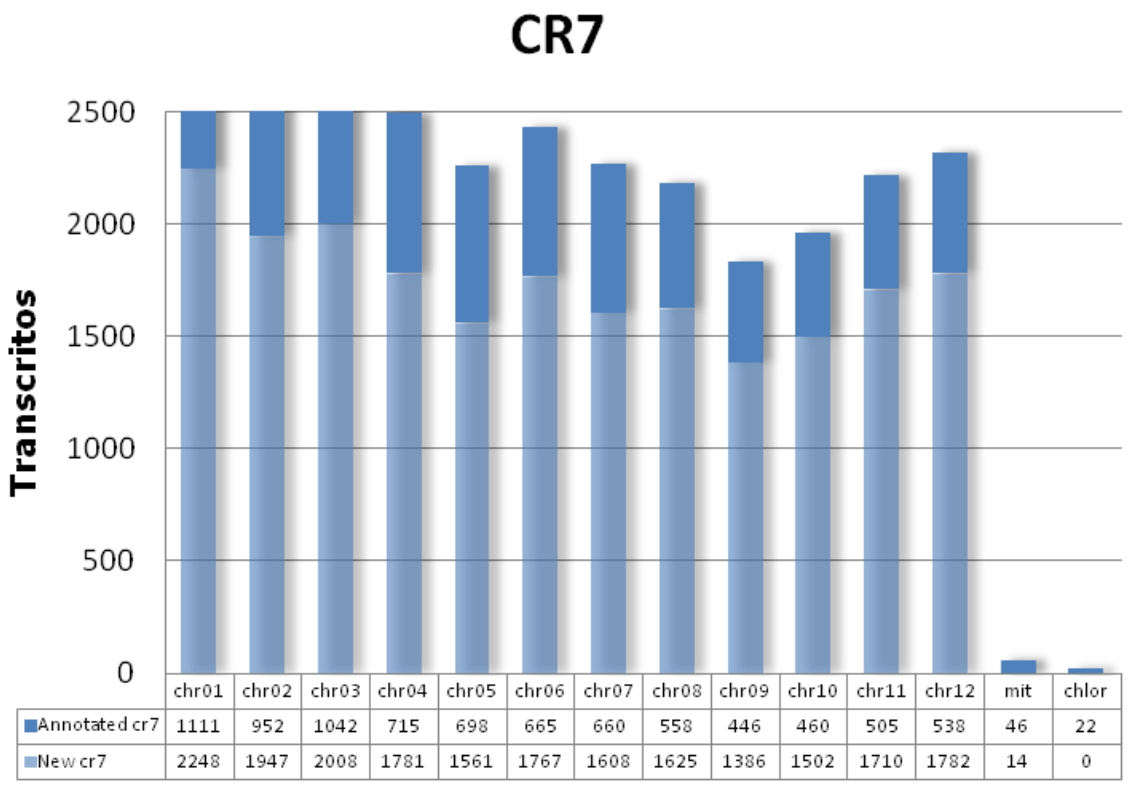
A altura das barras claras representa os transcritos novos enquanto a altura das barras escuras representa os genes anotados. Logo abaixo das barras, apresenta-se a tabela de dados.



	chr01	chr02	chr03	chr04	chr05	chr06	chr07	chr08	chr09	chr10	chr11	chr12	mit	chlor
Annotated ir3	833	721	774	505	521	498	463	413	307	316	362	394	47	39
New ir3	937	771	868	691	595	719	707	608	513	612	692	747	8	0

Figura 58 - Relação entre transcritos anotados e transcritos novos para as bibliotecas inoculadas de 3 dias, por cromossomo.

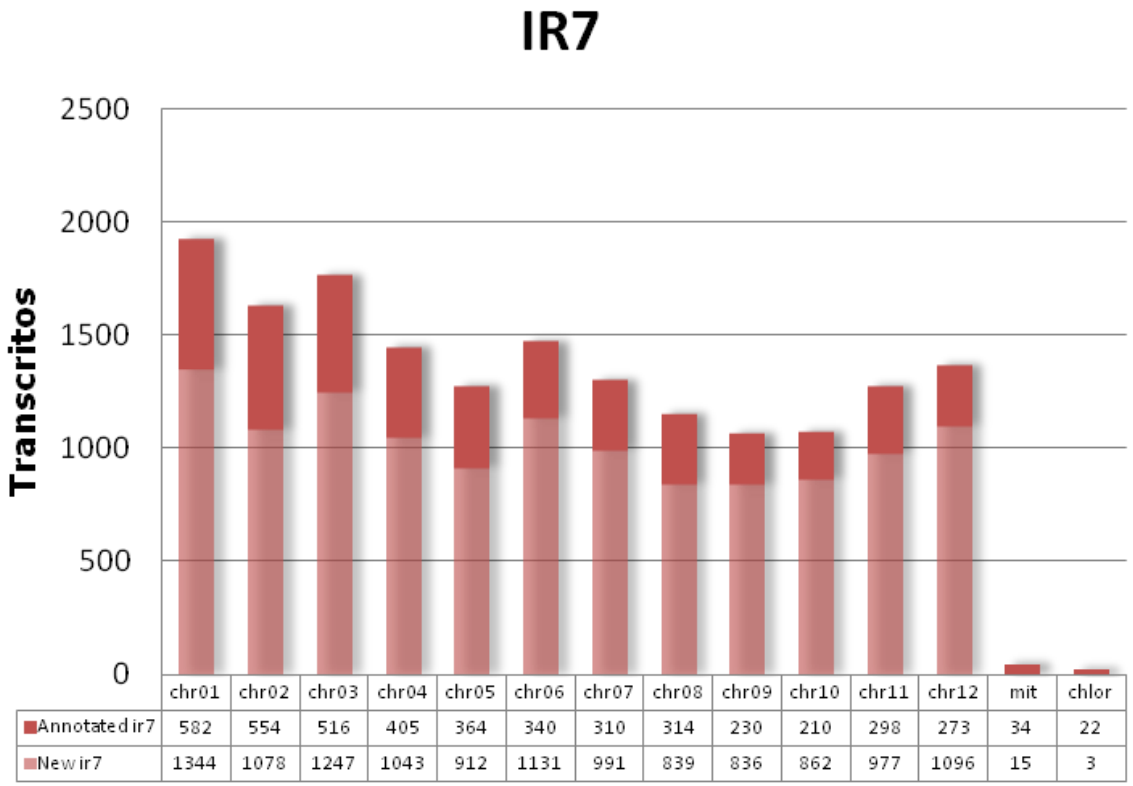
A altura das barras claras representa os transcritos novos enquanto a altura das barras escuras representa os genes anotados. Logo abaixo das barras, apresenta-se a tabela de dados.



	chr01	chr02	chr03	chr04	chr05	chr06	chr07	chr08	chr09	chr10	chr11	chr12	mit	chlor
Annotated cr7	1111	952	1042	715	698	665	660	558	446	460	505	538	46	22
New cr7	2248	1947	2008	1781	1561	1767	1608	1625	1386	1502	1710	1782	14	0

Figura 60 - Relação entre transcritos anotados e transcritos novos para as bibliotecas controle de 7 dias, por cromossomo.

A altura das barras claras representa os transcritos novos enquanto a altura das barras escuras representa os genes anotados. Logo abaixo das barras, apresenta-se a tabela de dados.



	chr01	chr02	chr03	chr04	chr05	chr06	chr07	chr08	chr09	chr10	chr11	chr12	mit	chlor
Annotated ir7	582	554	516	405	364	340	310	314	230	210	298	273	34	22
New ir7	1344	1078	1247	1043	912	1131	991	839	836	862	977	1096	15	3

Figura 59 - Relação entre transcritos anotados e transcritos novos para as bibliotecas inoculadas de 7 dias.

A altura das barras claras representa os transcritos novos enquanto a altura das barras escuras representa os genes anotados. Logo abaixo das barras, apresenta-se a tabela de dados.

4.3. ANÁLISE DOS *SPLICINGS* ALTERNATIVOS

Um dos itens dos arquivos de saída do pipeline Cufflinks é a informação sobre o *locus* em que cada transcrito montado foi mapeado contra o genoma de referência (Figura 61). Esta informação foi utilizada para verificar a correspondência do trancrito mapeado contra um gene da referência com relação à estrutura dos éxons. Sempre que um transcrito mapeado contra um gene apresentou uma variação de pelo menos uma base *upstream* ou *downstream* para qualquer éxon, este foi considerado como *splicing* alternativo. Esta informação foi utilizada para verificar quais genes do genoma do arroz sofreram *splicing* alternativo nos transcritos encontrados nas bibliotecas analisadas. A Figura 62 apresenta 4 gráficos com esta distribuição. É possível notar que em todos os cromossomos houve um percentual de *splicing* alternativo em quase 100% dos genes anotadores. Ou seja, ocorre o evento de *splicing* alternativo em quase 100% dos genes expressos em arroz na condição de inoculado com a bactéria *H. seropedicae*. Estes são chamados genes anotadores, porque não corresponde a todos os genes anotados do genoma do arroz, mas sim aos expressos nesta condição e que correspondem a um determinado transcrito.

Comparando com os dados da Figura 62, os gráficos das Figuras 63 e 64 mostram dos trancritos obtidos, qual percentual sofre *splicing* alternativo. Enquanto a Figura 62 mostra dos genes anotadores, qual percentual está sujeito a *splicing*.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2(fold_change)	test_stat	p_value	q_value	significant	
2	TSS1	XLOC_000001	ndhA	chloroplast:109999-113889	cr3	ir3	NOTEST	0	222.475	1.79769e+308	1.79769e-	0.0908337	1	no	
3	TSS10	XLOC_000010	Os01t0105300-01	chromosome01:289623-290002	cr3	ir3	OK	601.297	339.581	-0.824321	0.712376	0.476232	0.697324	no	
4	TSS100	XLOC_000094	-	chromosome01:7213231-7213254	cr3	ir3	OK	0	10499.2	1.79769e+308	1.79769e-	0.0079439	0.0891053	no	
5	TSS1000	XLOC_000936	Os01t0764600-02	chromosome01:33915042-33919437	cr3	ir3	OK	468.412	322.315	-0.539305	0.807045	0.419641	0.661237	no	
6	TSS1001	XLOC_000937	Os01t0765000-01	chromosome01:33962821-33965067	cr3	ir3	OK	29.357	174.234	-0.75268	0.820707	0.411813	0.657644	no	
7	TSS1002	XLOC_000938	Os01t0766900-01	chromosome01:34048253-34050772	cr3	ir3	OK	170.958	152.726	-0.162699	0.325977	0.744442	0.861307	no	
8	TSS1003	XLOC_000939	Os01t0766966-00	chromosome01:34050835-34051679	cr3	ir3	OK	448.929	294.46	-0.608417	0.616205	0.537759	0.73682	no	
9	TSS1004	XLOC_000940	-	chromosome01:34058564-34061330	cr3	ir3	LOWDATA	0	0	-1.79769e+308	0	1	1	no	
10	TSS1005	XLOC_000940	-	chromosome01:34058564-34061330	cr3	ir3	OK	364.146	584.146	-264.012	0.674434	0.500035	0.713538	no	
11	TSS1006	XLOC_000940	-	chromosome01:34058564-34061330	cr3	ir3	OK	242.98	152.716	-0.669988	133.364	0.182321	0.477227	no	
12	TSS1007	XLOC_000941	Os01t0771200-01	chromosome01:34274265-34274315	cr3	ir3	OK	0	576.942	1.79769e+308	1.79769e-	0.0507912	0.250655	no	
13	TSS1008	XLOC_000942	Os01t0771300-01	chromosome01:34275769-34282272	cr3	ir3	OK	7605.61	0	-1.79769e+308	-1.79769e	0.0002371	0.0086436	yes	
14	TSS1009	XLOC_000943	Os01t0771400-00	chromosome01:34282766-34297096	cr3	ir3	FAIL	117.317	125.876	0	0	1	1	no	
15	TSS101	XLOC_000095	Os01t0230200-01	chromosome01:7215512-7228163	cr3	ir3	OK	2412.63	0	-1.79769e+308	-1.79769e	0.0446755	0.237028	no	
16	TSS1010	XLOC_000943	Os01t0771400-00	chromosome01:34282766-34297096	cr3	ir3	FAIL	225.25	195.651	0	0	1	1	no	
17	TSS1011	XLOC_000944	Os01t0771350-01	chromosome01:34282766-34297096	cr3	ir3	FAIL	734.838	140.979	0	0	1	1	no	
18	TSS1012	XLOC_000945	Os01t0772000-01	chromosome01:34316595-34321556	cr3	ir3	OK	464.386	241.458	-0.943554	11.826	0.236966	0.529761	no	
19	TSS1013	XLOC_000946	Os01t0772200-01	chromosome01:34327764-34331485	cr3	ir3	OK	919.033	645.489	-0.509723	0.777895	0.436631	0.671358	no	
20	TSS1014	XLOC_000947	Os01t0773200-02	chromosome01:34404746-34411088	cr3	ir3	OK	666.533	388.639	-0.778245	160.366	0.108789	0.376972	no	
21	TSS1015	XLOC_000948	Os01t0775100-01	chromosome01:34488837-34494434	cr3	ir3	OK	145.362	869.635	-0.741164	145.237	0.146398	0.432406	no	
22	TSS1016	XLOC_000949	-	chromosome01:34496210-34496230	cr3	ir3	OK	3032.46	14034	221.036	-125.032	0.211183	0.508279	no	
23	TSS1017	XLOC_000950	Os01t0775300-01	chromosome01:34500693-34513252	cr3	ir3	OK	128.365	749.966	-0.775351	130.989	0.190235	0.48653	no	
24	TSS1018	XLOC_000951	Os01t0776500-01	chromosome01:34573235-34573255	cr3	ir3	OK	14718.2	50332.8	17.739	-177.882	0.0752701	0.310867	no	
25	TSS1019	XLOC_000952	Os01t0776700-01	chromosome01:34584896-34587778	cr3	ir3	OK	121.971	482.961	-133.656	190.825	0.0563594	0.264487	no	
26	TSS102	XLOC_000096	-	chromosome01:7258567-7258613	cr3	ir3	OK	247.576	700.534	150.058	-110.247	0.270258	0.560445	no	
27	TSS1020	XLOC_000953	Os01t0779400-02	chromosome01:34729758-34739902	cr3	ir3	FAIL	56.534	441.054	0	0	1	1	no	
28	TSS1021	XLOC_000953	Os01t0779400-02	chromosome01:34729758-34739902	cr3	ir3	FAIL	415.064	291.882	0	0	1	1	no	
29	TSS1022	XLOC_000954	Os01t0783200-02	chromosome01:34915140-34922660	cr3	ir3	OK	308.495	15.427	-0.999793	13.693	0.170906	0.464463	no	
30	TSS1023	XLOC_000955	Os01t0795300-01	chromosome01:35420433-35426201	cr3	ir3	NOTEST	145.309	133.817	-0.118865	0.131475	0.8954	1	no	
31	TSS1024	XLOC_000956	Os01t0795700-01	chromosome01:35433597-35433647	cr3	ir3	OK	958.349	########	105.496	-0.990215	0.322069	0.598779	no	

Figura 61 - Uma das tabelas de saída do Cufflinks.

Possui identificação do teste_id como TSS e *locus* em que cada transcrito montado foi mapeado contra o genoma de referência. Esta informação foi utilizada para verificar a correspondência do trancrito mapeado contra um gene da referência com relação à estrutura dos éxons.

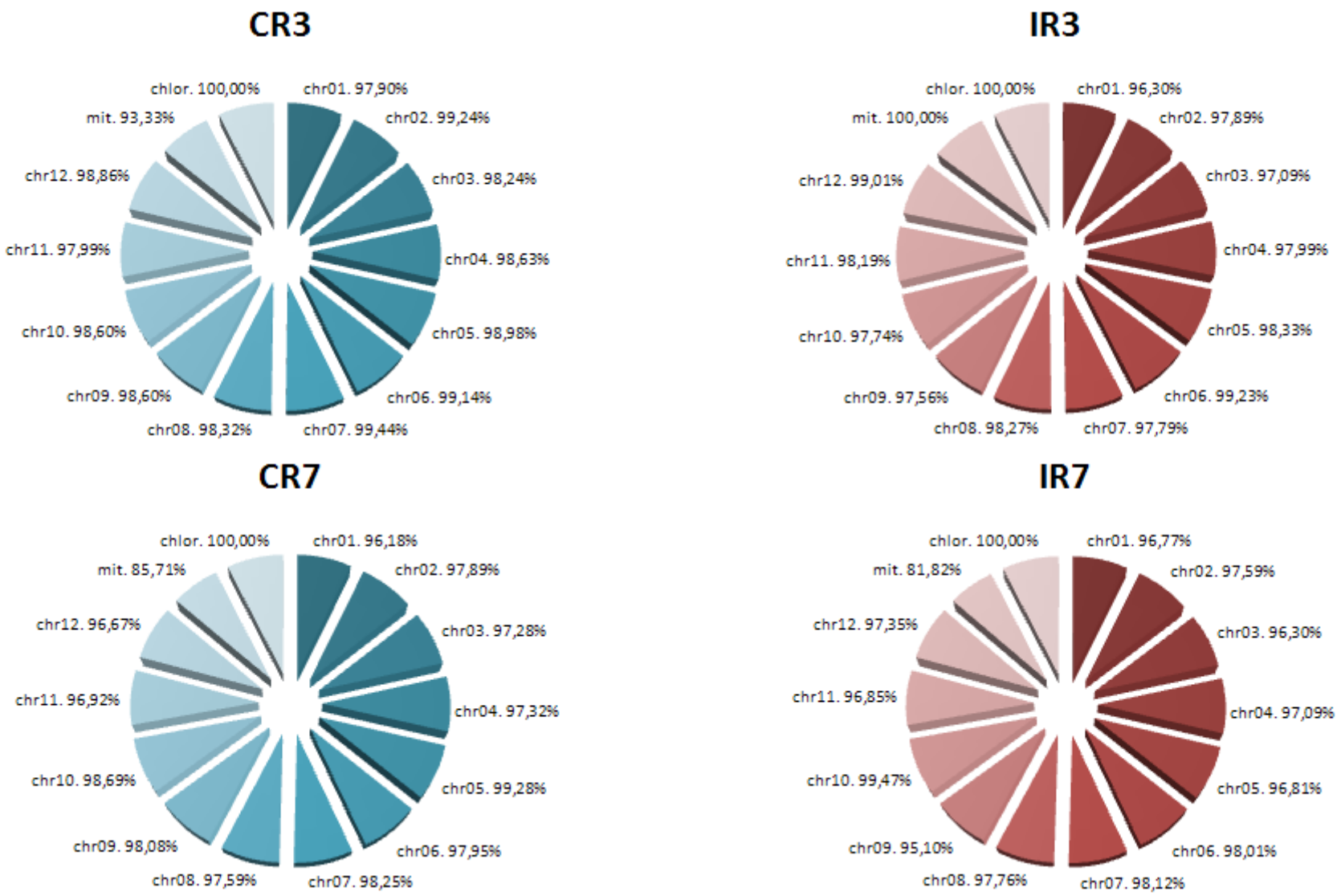


Figura 62 – Distribuição por cromossomo do percentual de genes anotadores sujeitos a *splicing* alternativo.

Possível notar que em todos os cromossomos houve um percentual de *splicing* alternativo em quase 100% dos genes anotadores.

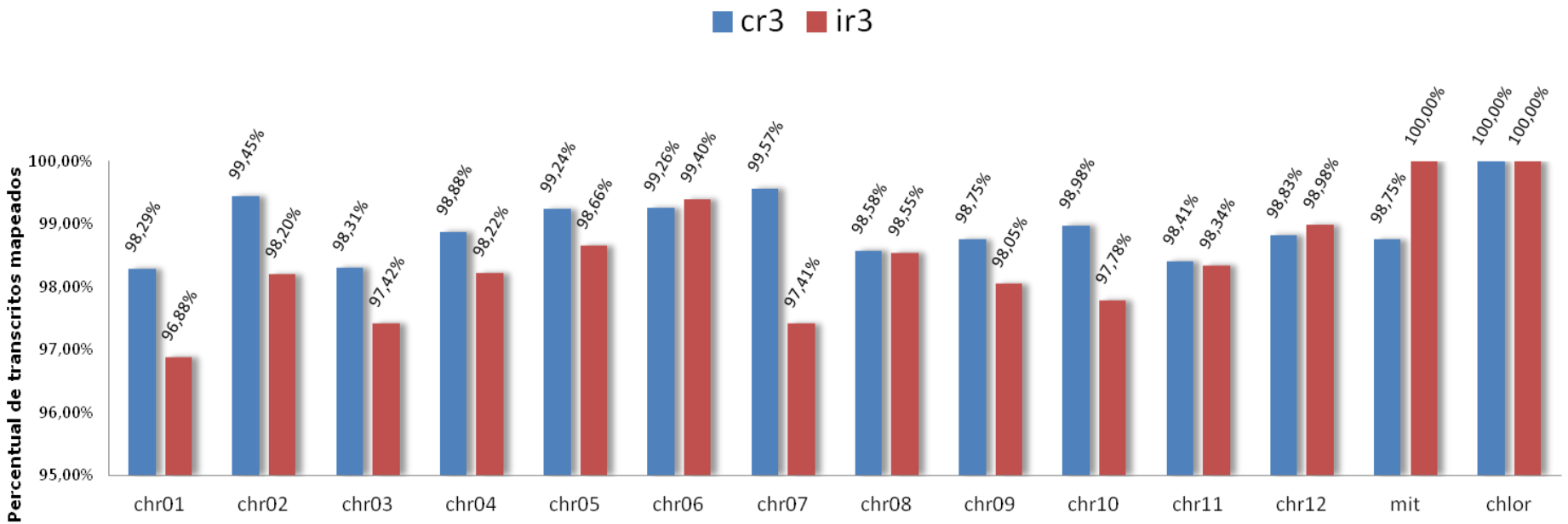


Figura 63 - Distribuição por cromossomo do percentual de transcritos sujeitos a *splicing* alternativo nas bibliotecas de 3 dias controle e inoculado.

Mostra dos transcritos obtidos, qual percentual sofre *splicing* alternativo.

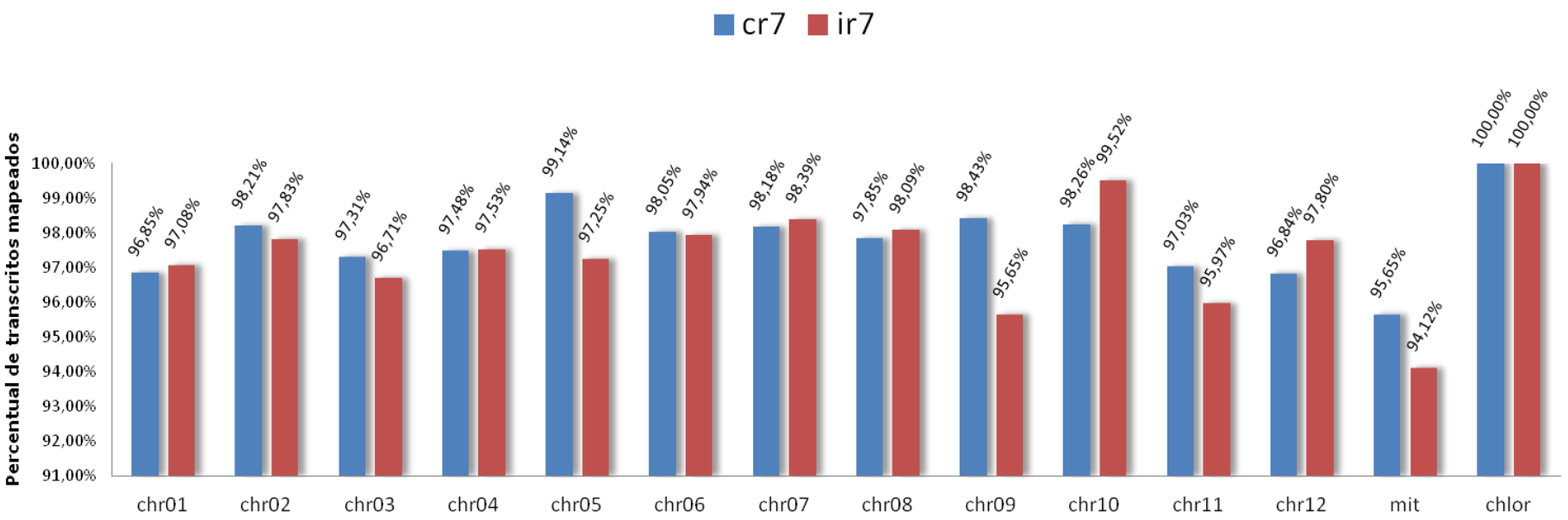


Figura 64 - Distribuição por cromossomo do percentual de transcritos sujeitos a splicing alternativo nas bibliotecas de 7 dias controle e inoculada.

Mostra dos transcritos obtidos, qual percentual sofre *splicing* alternativo.

4.4. DADOS DIFERENCIALMENTE EXPRESSOS

As análises apresentadas até o momento foram realizadas com os arquivos de saída do Cuffmerge, que apresenta a sobreposição das unidades de trancrição obtidas (CUFF) para cada biblioteca montando os transcritos considerados até o momento (TCONS), bem como o mapeamento destes transcritos contra o genoma de referência do arroz (uma referência para cada cromossomo e organelas). As sessões seguintes, que tratam os dados diferencialmente expressos, utiliza os arquivos de saída do programa Cuffdiff.

4.4.1. Resultados de 3 dias após inoculação (CR3 vs. IR3).

A Tabela 12 apresenta os resultados da expressão diferencial para 3 dias após inoculação, e os gráficos das Figuras 65-67 ilustram esses dados. Foram encontrados 21.539 TSS, sendo 2.248 D.E. Além de um total de 10.536 transcritos novos. Para estes dados, foram comparadas as condições controle *versus* inoculado para o tratamento de 3 dias (CR3 vs. IR3).

Tabela 12 - Contagem dos dados Cuffdiff para 3 dias após inoculação (CR3 vs. IR3).

Cromossomos	Genes	TSS	TSS DE	Transcritos novos
1	2.294	2.524	243	1110
2	1.955	2.184	229	956
3	2.130	2.388	147	981
4	1.682	1.844	218	939
5	1.517	1.681	226	775
6	1.622	1.794	98	888
7	1.543	1.674	157	854
8	1.445	1.561	55	811
9	1.115	1.217	244	658
10	1.239	1.352	190	747
11	1.438	1.542	204	902
12	1.502	1.654	170	911
mit	41	77	37	3
clor	25	47	30	1
TOTAL	**19548**	**21539**	**2248**	**10536**

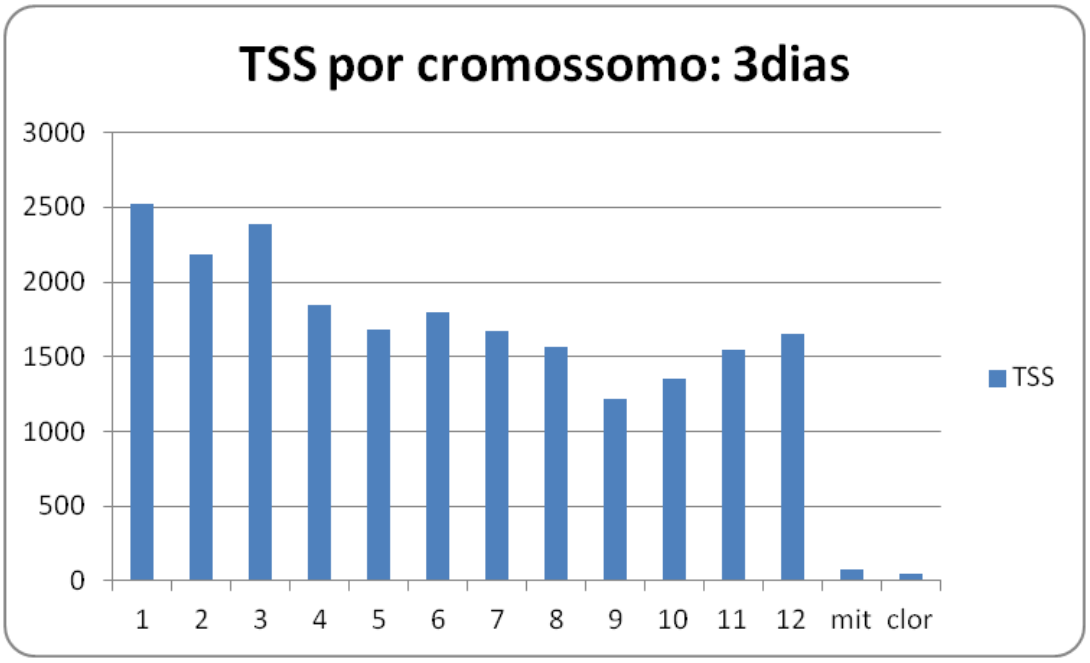


Figura 65 - Quantidade de TSS encontrados por cromossomo para as bibliotecas de 3 dias.

Os cromossomos 1 e 3 apresentam o maior número de TSS's.

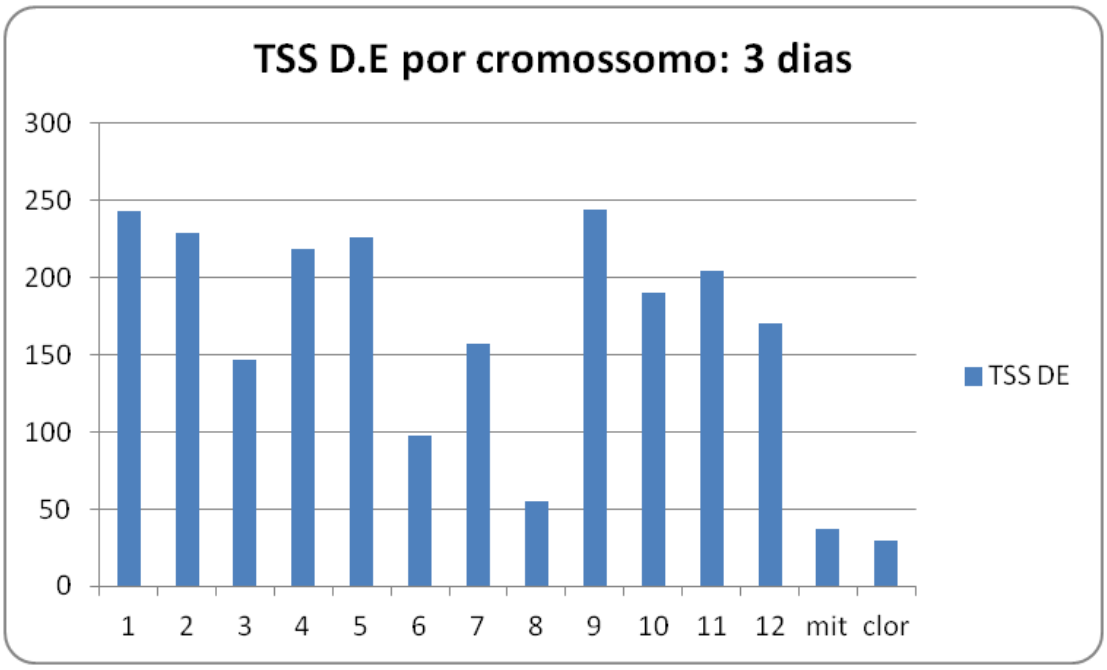


Figura 66 - Número de TSS's Diferencialmente Expressos por cromossomo para as bibliotecas de 3 dias.

Os cromossomos 1 e 9 apresentam os maiores resultados.

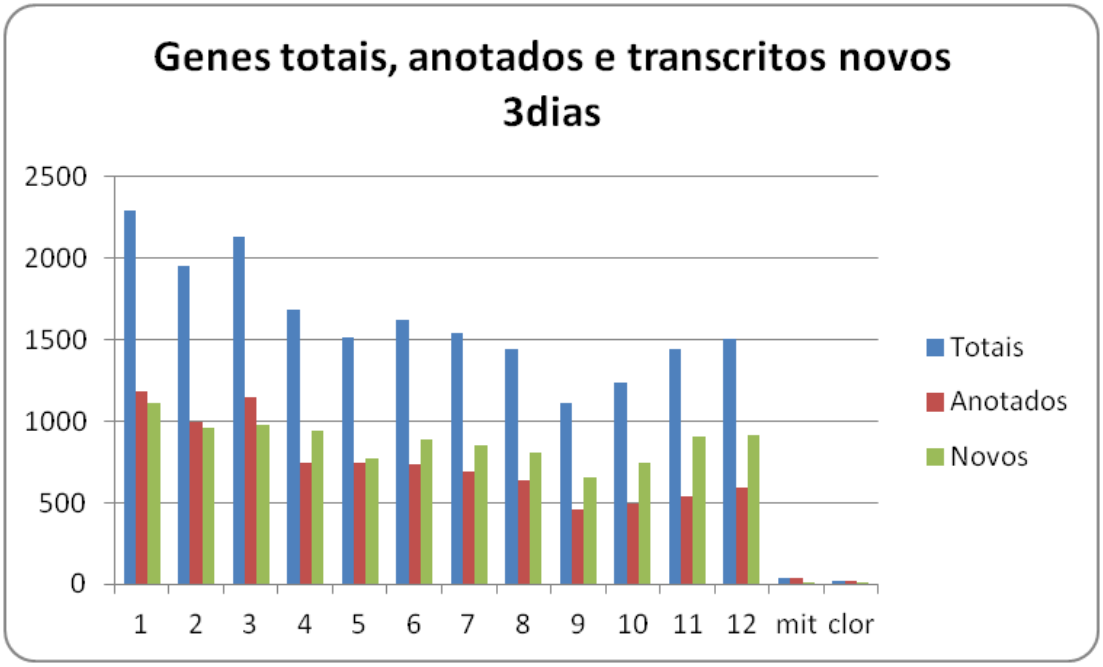


Figura 67 - Relação entre a quantidade de transcritos totais, ou seja, genes anotados e transcritos novos quantificados e em separado: Anotados e novos.

4.4.2. Resultados de 7 dias após inoculação (CR7 vs. IR7).

A Tabela 13 apresenta os resultados da expressão diferencial para 7 dias após inoculação, e as Figuras 68-70 apresentam os gráficos que ilustram esses dados. Foram encontrados 34.746 TSS, sendo 4.527 D.E., além de um total de 24.600 transcritos novos. Para estes dados, foram comparadas as condições controle versus inoculado para o tratamento de 7 dias (CR7 vs. IR7).

Tabela 13 - Contagem dos dados Cuffdiff para 7 dias após inoculação (CR7 vs. IR7).

Cromossomos	Genes	TSS	TSS DE	Transcritos novos
1	3.754	3.883	413	2583
2	3.299	3.440	519	2268
3	3.479	3.591	332	2343
4	2.920	2.994	713	2133
5	2.620	2.695	448	1865
6	2.823	2.899	433	2109
7	2.642	2.711	218	1951
8	2.479	2.537	141	1842
9	2.175	2.230	257	1675
10	2.231	2.300	385	1725
11	2.633	2.683	339	2051
12	2.609	2.689	260	2043
mit	39	58	44	11
clor	20	36	25	1
TOTAL	**33723**	**34746**	**4527**	**24600**

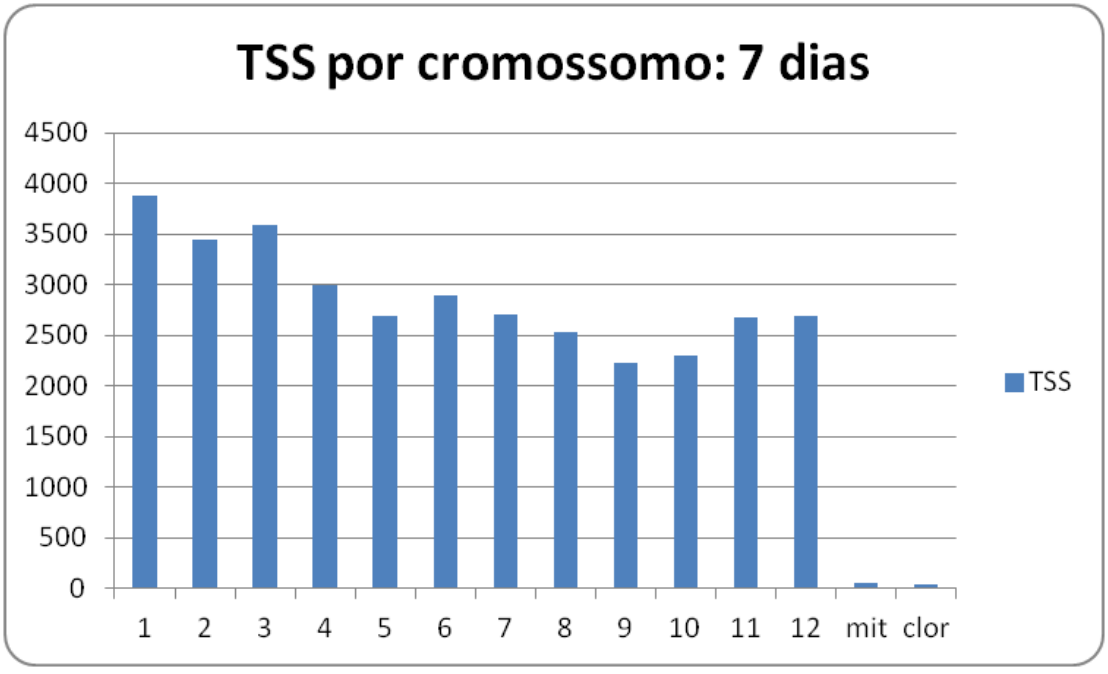


Figura 68 - Quantidade de TSS encontrados por cromossomo para as bibliotecas de 7 dias. Os cromossomos 1 e 3 apresentam os maiores resultados.

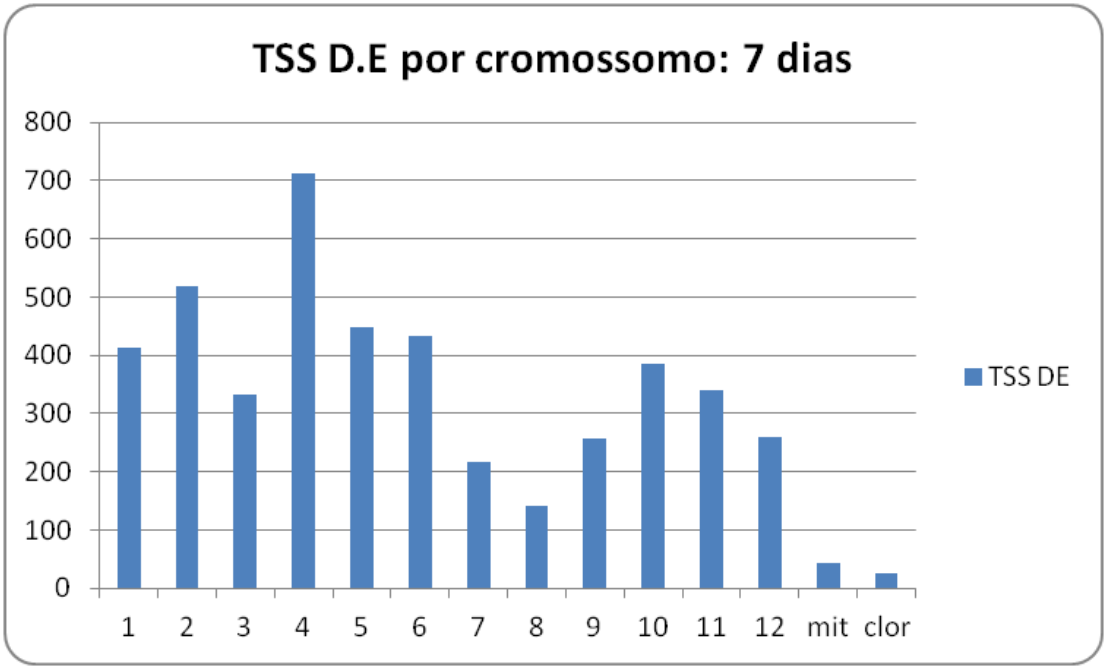


Figura 69 - Quantidade de TSS D.E encontrados por cromossomo para as bibliotecas de 7 dias.

O cromossomo 4 se destaca em maior número com TSS's D.E, posteriormente o cromossomo 2, que apresenta 519 TSS's D.E.

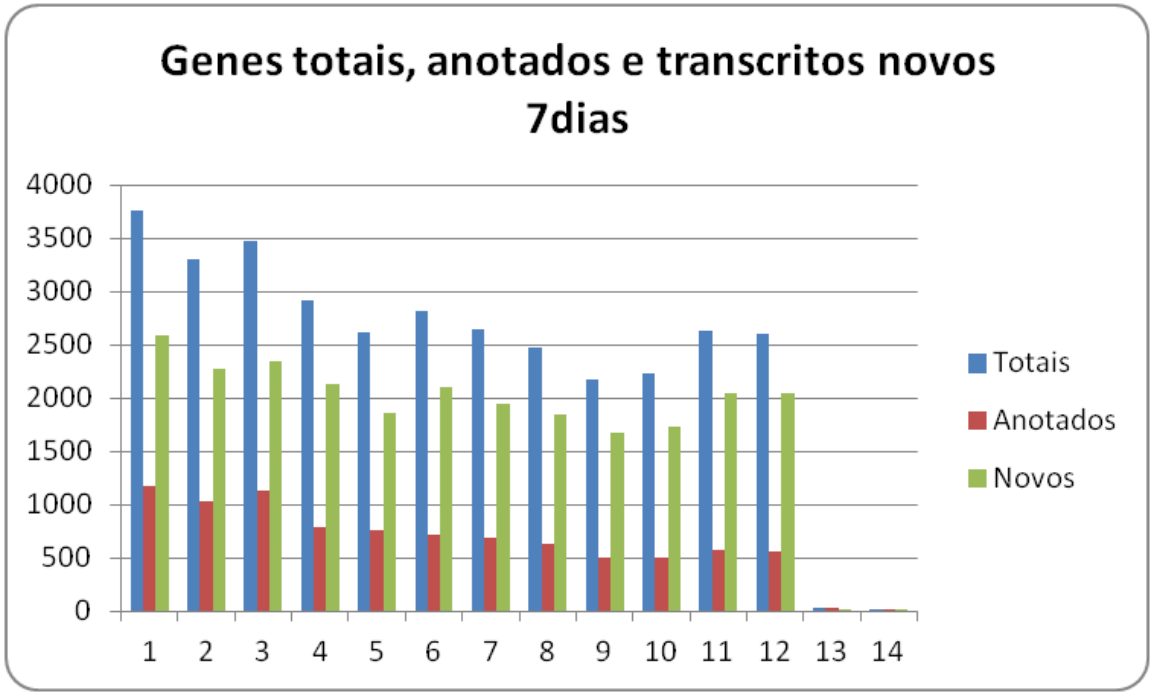


Figura 70 - Relação entre a quantidade de transcritos totais, ou seja, genes anotados e transcritos novos quantificados e em separado: Anotados e novos. Nota: 13=Mitocôndria, 14=Cloroplasto;

4.4.3. Resultados de 3 e 7 dias para índice único

Esta sessão descreve um experimento realizado para comparar a anotação e identificação de TSS's D.E. utilizando como índice de mapeamento todos os cromossomos constituindo um índice único do Bowtie/TopHat.

Para esta análise, todas as etapas descritas na sessão Materiais e Métodos são executadas, entretanto, como citado no paragrafo anterior, o índice do Bowtie é construído com todos os cromossomos juntos, dessa forma há a possibilidade de comparação com a utilização dos cromossomos constituíndo índices separados. O experimento indica que a diferença é muito grande quando comparado com a análise por cromossomo. O total de TSS de 3 dias somado nos cromossomos caiu de 21.539 para 8.527. Os TSS D.E. cai de 2.248 para 479, além dos transcritos novos que diminui 62,87%. Em 7 dias, o total de TSS passou de 34.746 para 14.909. Destes, os D.E. cairam de 4.527 para 1.523, além dos transcritos novos: de 24.600 para 10.432.

Uma possível explicação para essa diferença é o número de *matches* permitidos contra esta nova referência única. Nesta análise, cada *read* só pode alinhar 1 (uma) vez contra o novo índice (que nesta análise contém todos os cromossomos). Na análise por cromossomo ela poderia alinhar 1 vez em cada cromossomo, visto que cada cromossomo era tratado como um índice. A Tabela 14 apresenta os resultados obtidos neste experimento.

Tabela 14 - Resultado do alinhamento de índice único para 3 e 7 dias. CR x IR.

Cromossomo	Amostras	Genes	TSS	TSS DE	Transcritos novos
Todos	3 dias	7982	8527	479	3912
Todos	7dias	14635	14909	1523	10432

Para visualização da expressão dos transcritos é possível gerar um gráfico *Heatmap*, onde a lista da esquerda contêm os transcritos e as cores representam seu nível de expressão. Quanto mais escuro mais expresso. Na Figura 71, a coluna da esquerda refere-se a expressão verificada nas bibliotecas CR3 e a da direita nas bibliotecas IR3. Neste gráfico encontra-se exemplificada as variações de expressão entre os dois tratamentos de um total de 20 transcritos, quando realizada a análise

com o índice único. Gráficos como este podem ser gerados para quaisquer conjuntos de transcritos que se deseje.

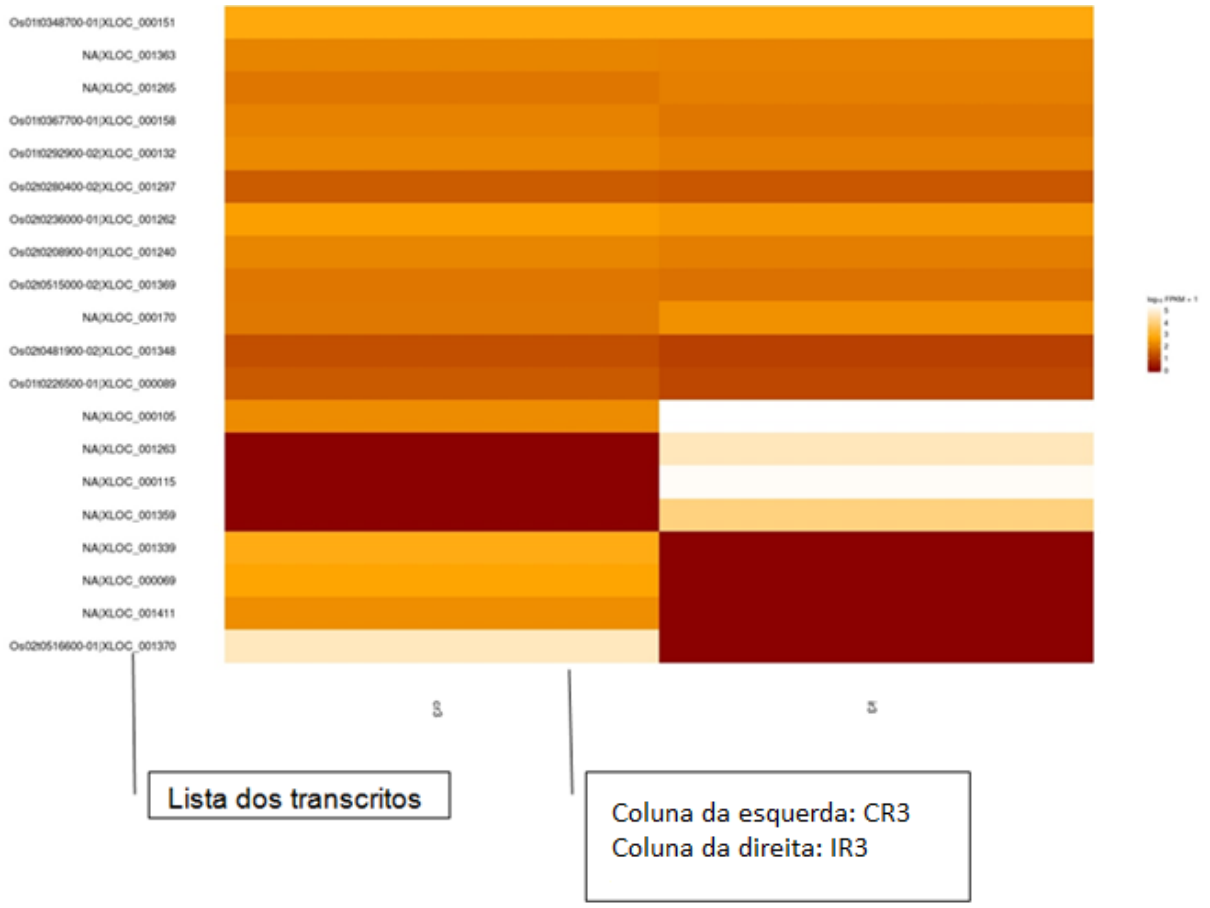


Figura 71 - Nível de expressão de 20 transcritos de índice único.

A coluna da esquerda refere-se a expressão verificada nas bibliotecas CR3 e a da direita nas bibliotecas IR3. Os valores estão mostrados em Log de RPKM + 1. Possível verificar que um determinado transcrito tem um nível de expressão diferente dependendo da biblioteca analisada.Gráfico gerado com a biblioteca CummeRbund do R Bioconductor, e mesmo alterando os valores de composição do gráfico, não foi possível aumentar a resolução da imagem.

4.4.4. Gráficos gerados para melhor visualização e interpretação dos dados

Após as análises apresentadas um grande número de gráficos podem ser gerados com o pacote CummeRbund do R/Bioconductor. As Figuras 72 e 74 apresentam os mapas de densidade gênica[8] para cada cromossomo, permitindo a comparação dos padrões observados entre cromossomos e entre tratamentos. Com exceção da sequências mapeadas contra a referência da mitocôndria e do cloroplasto, que são pouco representativos, todos os outros gráficos se mostram bastante relacionados. Para exemplificar de maneira mais clara, a Figura 73 apresenta 2 gráficos referentes ao cromossomo 10 com os dados plotados para 3 e 7 dias das bibliotecas controle *versus* inoculado respectivamente.

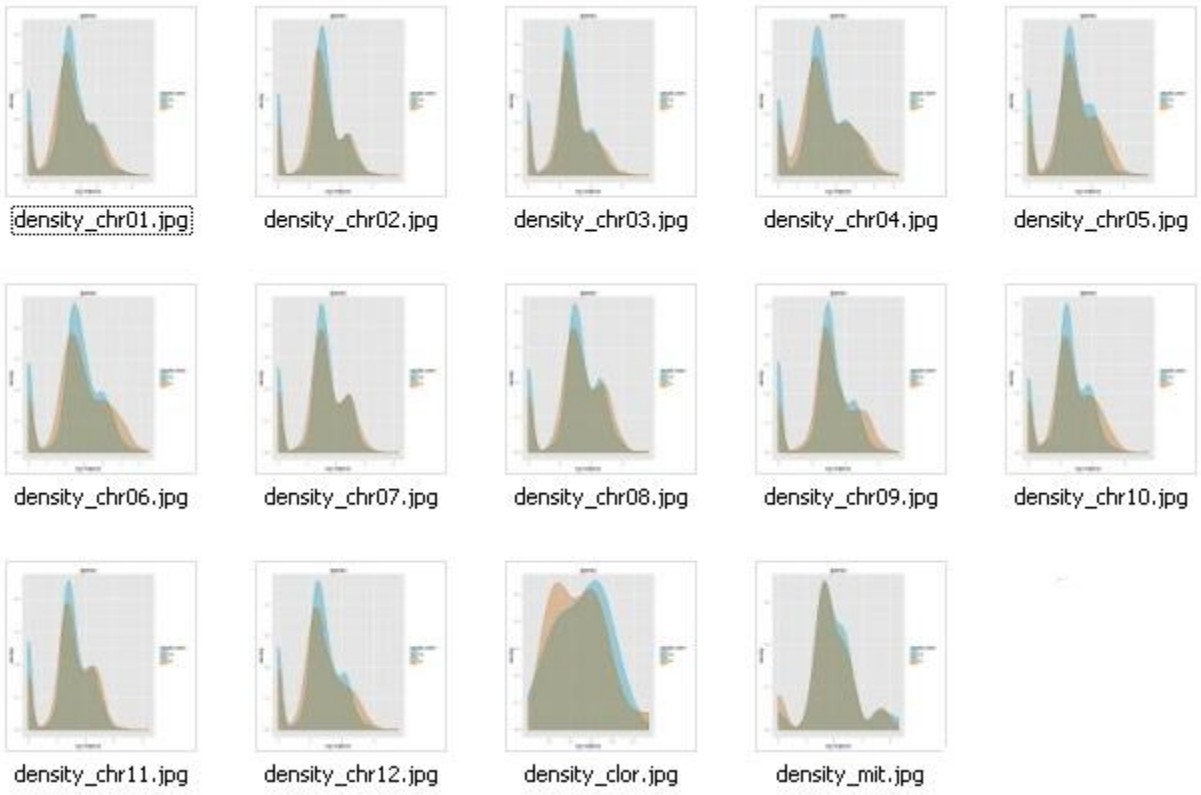


Figura 72 - Mapas de densidade gênica das amostras de 3 dias por cromossomo permitindo a comparação dos padrões observados entre cromossomos e entre tratamentos.

Com exceção da sequências mapeadas contra a referência da mitocôndria e do cloroplasto, que são pouco representativos, todos os outros gráficos se mostram bastante relacionados.

[8] Distribuição do nível de expressão para cada amostra.

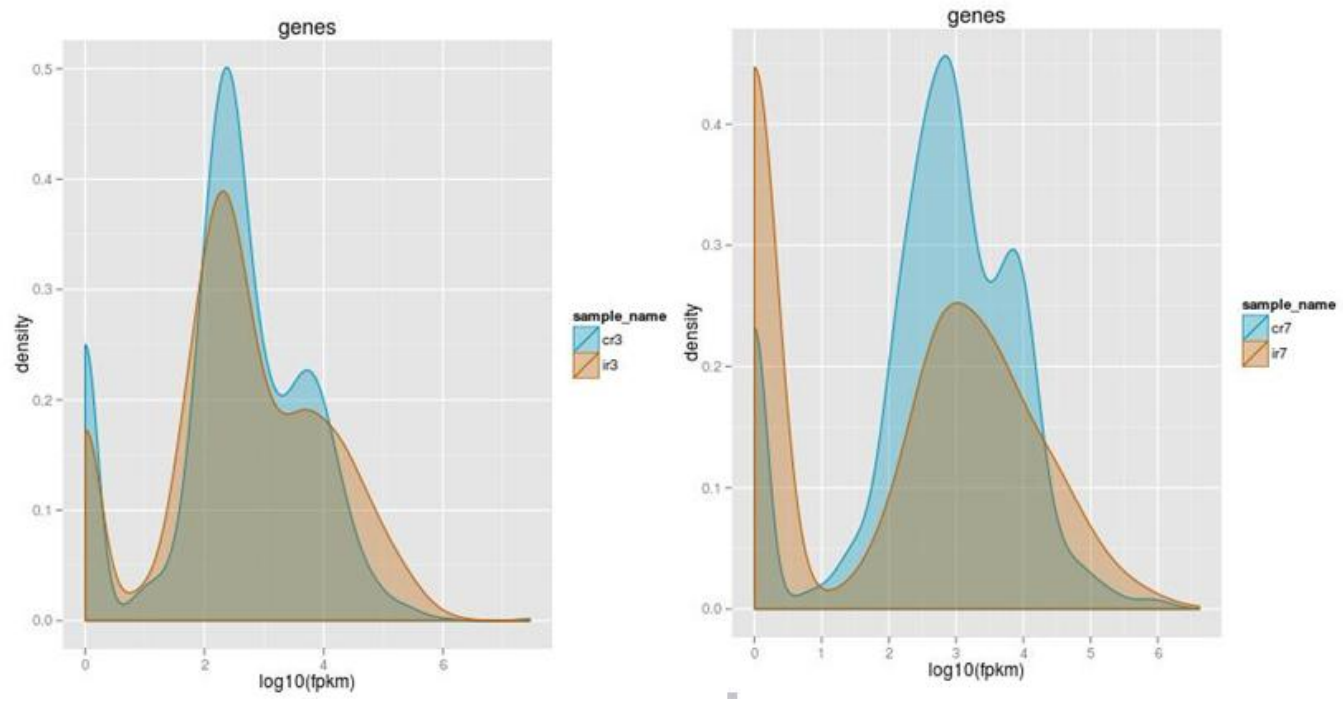


Figura 74 – Detalhamento do mapa de densidade gênica obtido para o cromossomo 10 para as bibliotecas de tratamento e controle de 3 dias (esquerda) e 7 dias (direita).

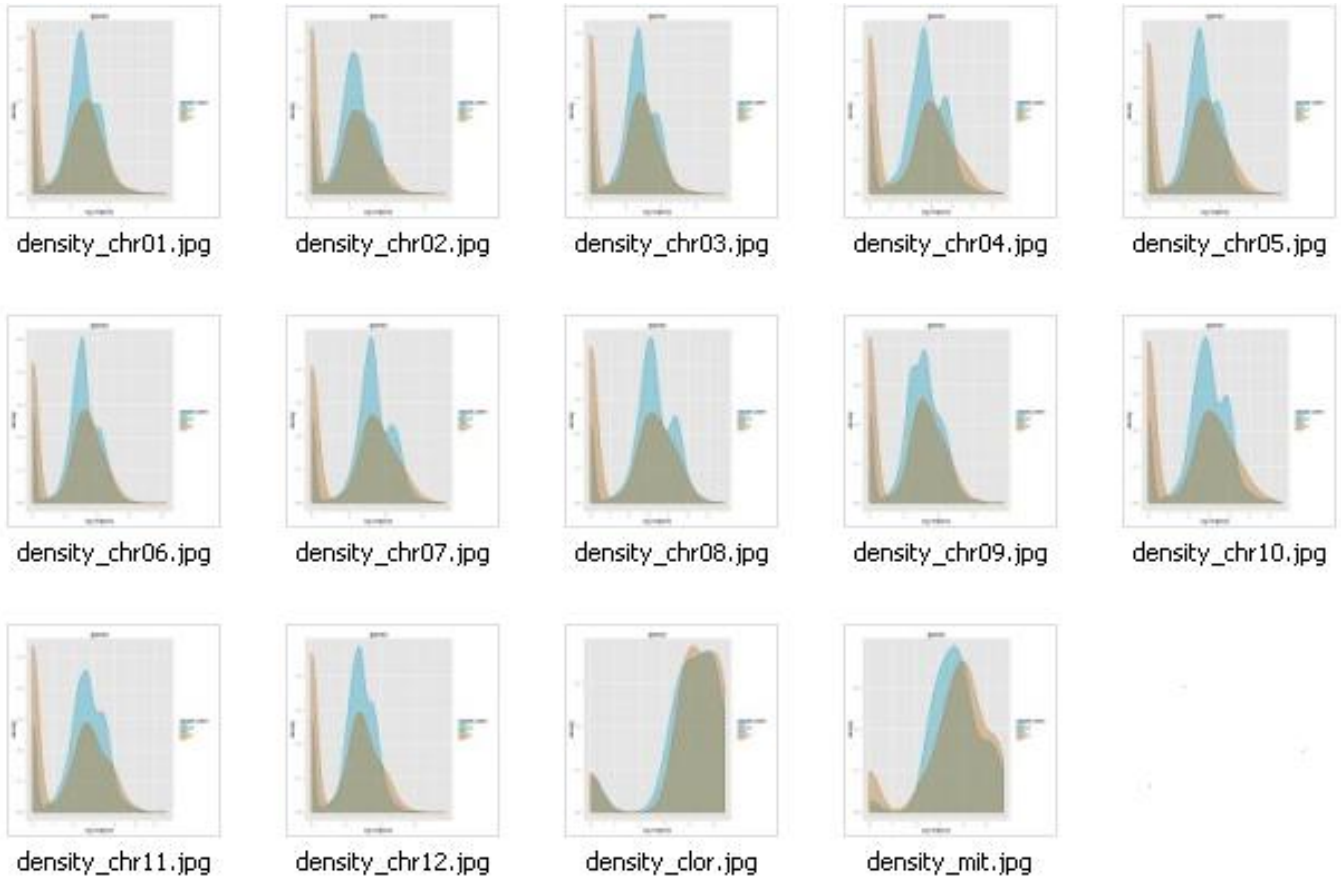


Figura 73 - Mapas de densidade gênica das amostras de 7 dias por cromossomo permitindo a comparação dos padrões observados entre cromossomos e entre tratamentos.

Com exceção da sequências mapeadas contra a referência da mitocôndria e do cloroplasto, que são pouco representativos, todos os outros gráficos se mostram bastante relacionados.

A Figura 75 apresenta um gráfico de dispersão (*scatter plot*) comparando a expressão dos TSS's entre as condições controle e inoculado para 3 dias (CR3 vs. IR3) considerando a análise de índice único. Semelhantemente, a Figura 76 apresenta um gráfico do tipo volcano (*volcano plot*) contendo os TSS's diferencialmente expressos (em azul) quando comparadas as taxas de expressão entre as condições controle e inoculada para 7 dias (CR7 vs. IR7)

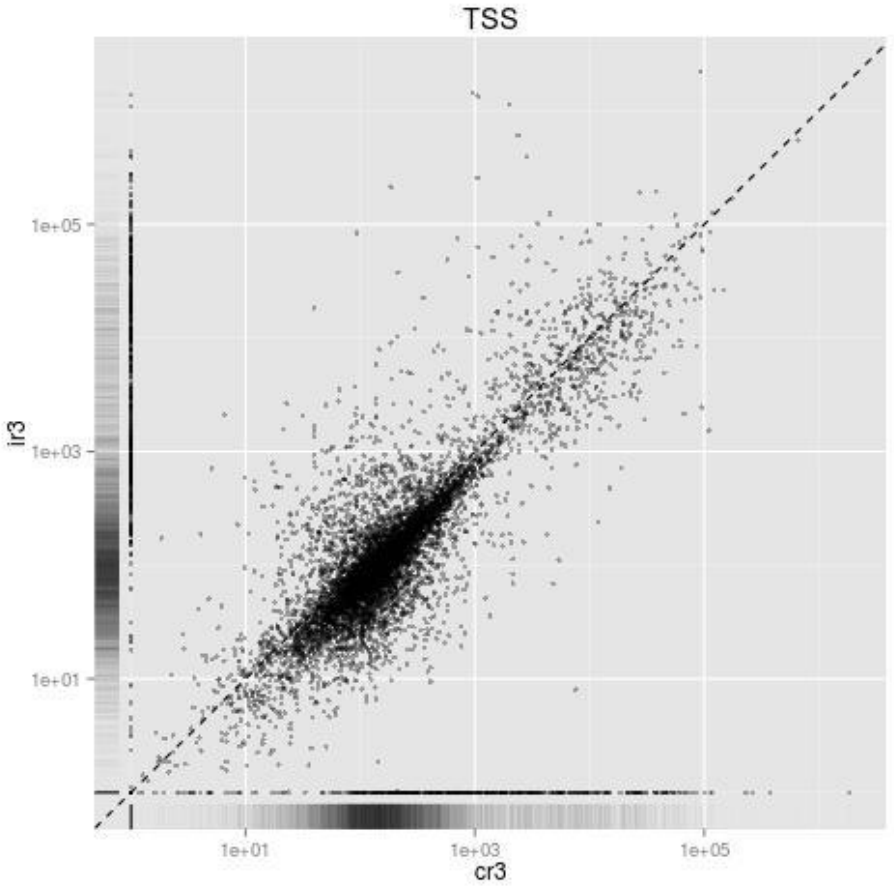


Figura 75 - Visualização da dispersão dos dados de TSS nas bibliotecas de 3 dias considerando os dados de índice único.

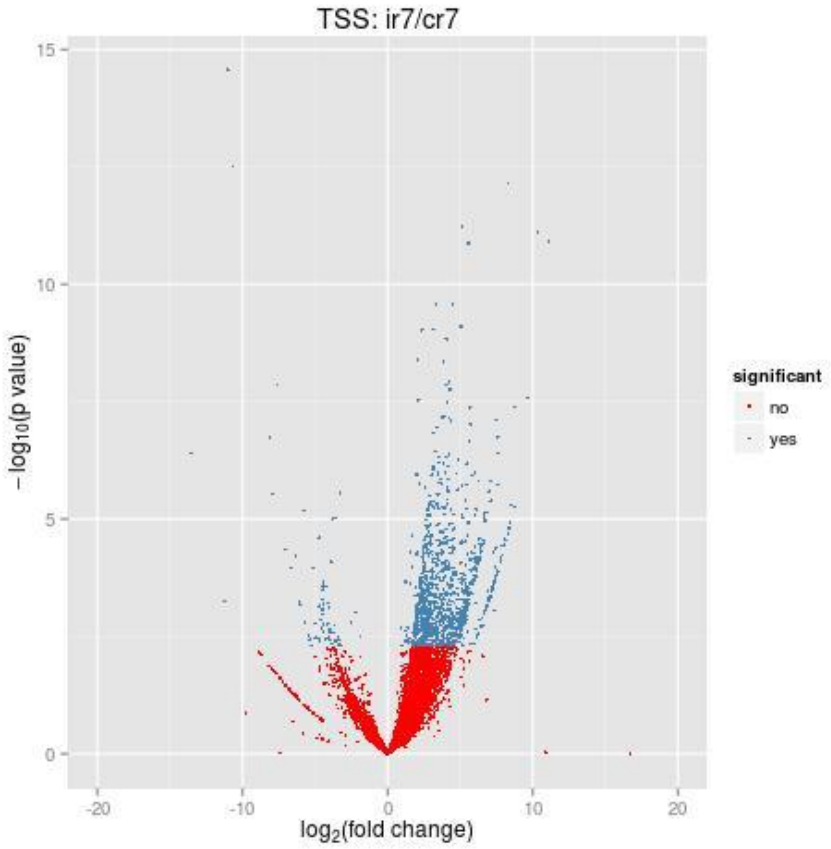


Figura 76 - TSS D.E (em azul) quando comparadas as taxas de expressão entre as condições controle e inoculado para 7 dias (CR7 vs. IR7).

Também foi avaliada a expressão dos transcritos em cada uma das replicatas de um determinado tratamento conforme exemplificado pela Figura 77. Este gráfico mostra a expressão do transcrito XLOC_000151 nas replicatas controle (CR3A e CR3B) e inoculadas (IR3A e IR3B). O gráfico foi gerado em R, para facilitar a visualização (devido às pequenas fontes criadas por *default*) os valores da abscissa são (CR3A, CR3B, IR3A e IR3B) e a ordenada refere-se ao FPKM calculado pelo Cufflinks encontrado em cada replicata.

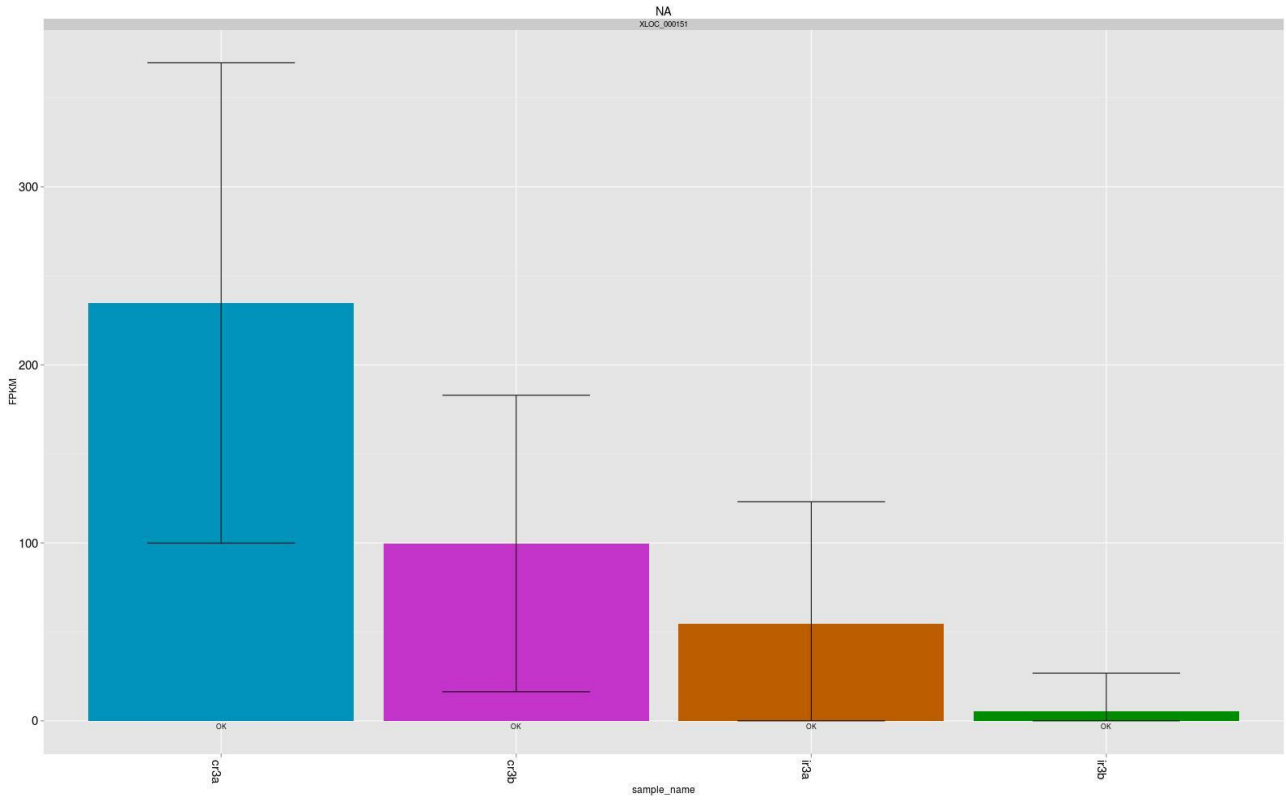


Figura 77 – Nível de expressão do transcrito XLOC_000151 nas replicatas controle (CR3A e CR3B) e inoculadas (IR3A e IR3B).

O gráfico foi gerado em R, para facilitar a visualização (devido às pequenas fontes criadas por *default*) os valores da abscissa são (CR3A, CR3B, IR3A e IR3B) e a ordenada refere-se ao FPKM calculado pelo Cufflinks encontrado em cada replicata.

5. CONCLUSÕES

Foi obtido um mapeamento médio total de 61,74% para *reads* sequenciadas do transcriptoma de arroz inoculado com a bactéria *H. seropedicae* contra o genoma de referência do arroz disponibilizado pelo RAP-DB.

As análises realizadas mostram que os resultados obtidos para as bibliotecas de 7 dias são, em média, o dobro daqueles obtidos para as bibliotecas de 3 dias. Isso corresponde também ao observado para a quantidade de transcritos novos encontrados. Na análise utilizando índice múltiplos (1 para cada cromossomo) foram encontrados 10.536 transcritos novos para as bibliotecas de 3 dias e 24.600 para as bibliotecas de 7 dias. Na análise utilizando o índice único (todos os cromossomos agrupados no mesmo índice) há 3.912 transcritos novos para as bibliotecas de 3 dias e 10.432 para as bibliotecas de 7 dias, ou seja, quase 3 vezes o valor encontrado de um tratamento com relação ao outro (7/3 dias). É possível concluir que após 7 dias de inoculação o arroz expressa uma maior quantidade de transcritos novos em resposta aos estímulos causados pela bactéria *H. seropedicae.*

Os mapas de densidade gênica seguem um padrão semelhante quando os cromossomos são comparados entre si. Talvez a diferença no padrão observado nas organelas (mitocôdria e cloroplasto) possa ser explicada pela quantidade reduzida de genes anotados bem como a quantidade de genes expressos nestas organelas. Neste trabalho, foi observado que pode haver uma variação da expressão de um determinado transcrito dentro das replicatas.

Os resultados variam quando a análise de expressão diferencial é realizada utilizando um índice múltiplo ou único. Um dos fatores contribuintes pode ser a utilização do limite de alinhamento, onde apenas um alinhamento da *read* por índice é permitido.

Um total de 6.942 (16,4%) e 4.915 (11,6%) genes obtiveram cobertura nas bibliotecas CR3 e IR3; e 6.733 (15,9%) e 3.807 (9%) genes para as bibliotecas CR7 e IR7, respectivamente.

202.770.984 (61,7%) *reads* foram mapeadas no genoma do arroz: 216.980.165 (75,2%) são *reads* exônicas, e 71.493.386 (24,8%) mapearam em junções de éxons. Esses números totalizam 288.473.551 *reads* alinhadas, evidenciando que 85.702.567 dessas mapeavam em mais de um cromossomo em separado.

Todos os cromossomos apresentam um percentual superior a 96% de *splicing* alternativo nos genes anotadores, característica essa observada nos transcritos que mapearam nas regiões correspondentes à estes genes no genoma de referência.

6. PERSPECTIVAS

Algumas perguntas ainda puderam ser levantadas ao final das análises apresentadas neste trabalho, por exemplo: o que representam os transcritos novos identificados? Uma abordagem para obter a resposta à esta pergunta, foi a realização de um alinhamento de todos os transcritos novos dos arquivos de saída do Cuffmerge contra o banco de dados NR (non-redundant) do NCBI Protein, conforme o gráfico da Figura 78. Após o alinhamento com o BLASTP, 691 (3,31%) transcritos identificados como novos no genoma do arroz obtiveram alinhamento (*match*) com sequências de proteínas depositadas no NR. Para análises futuras, os 20.185 (96,7%) transcritos novos restantes poderão ser comparados à um banco de dados de microRNA, visto que, a estrutura secundária de vários destes, condizem com a estrutura de precursores de microRNA (LOPES, COVRE, *et al.*, 2012).

Outro ponto importante avaliado nesta etapa foi o tamanho dos transcritos novos encontrados. Diversos estudos mostraram a função regulatória de pequenos transcritos presentes na célula eucariótica conhecidos como micro-RNAs (KIDNER e MARTIENSSEN, 2004). Este trabalho não visa a identificação de tais moléculas, entretanto, para se ter uma idéia da distribuição do tamanho destes transcritos sem anotação, nós criamos o gráfico da Figura 79, onde é possível visualizar que 31% dos transcritos novos tem mais de 50 pb, tamanho compatível com algumas moléculas precursoras de micro-RNAs dupla fita (*pre-miRNA*) em eucariotos.

Realizamos a predição da estrutura secundária destas sequencias não-anotadas no genoma de arroz (LOPES, COVRE, *et al.*, 2012), e pudemos verificar que algumas estruturas obtidas estavam em acordo com as estruturas depositadas em bancos de dados como o miRBase (Figura 80).

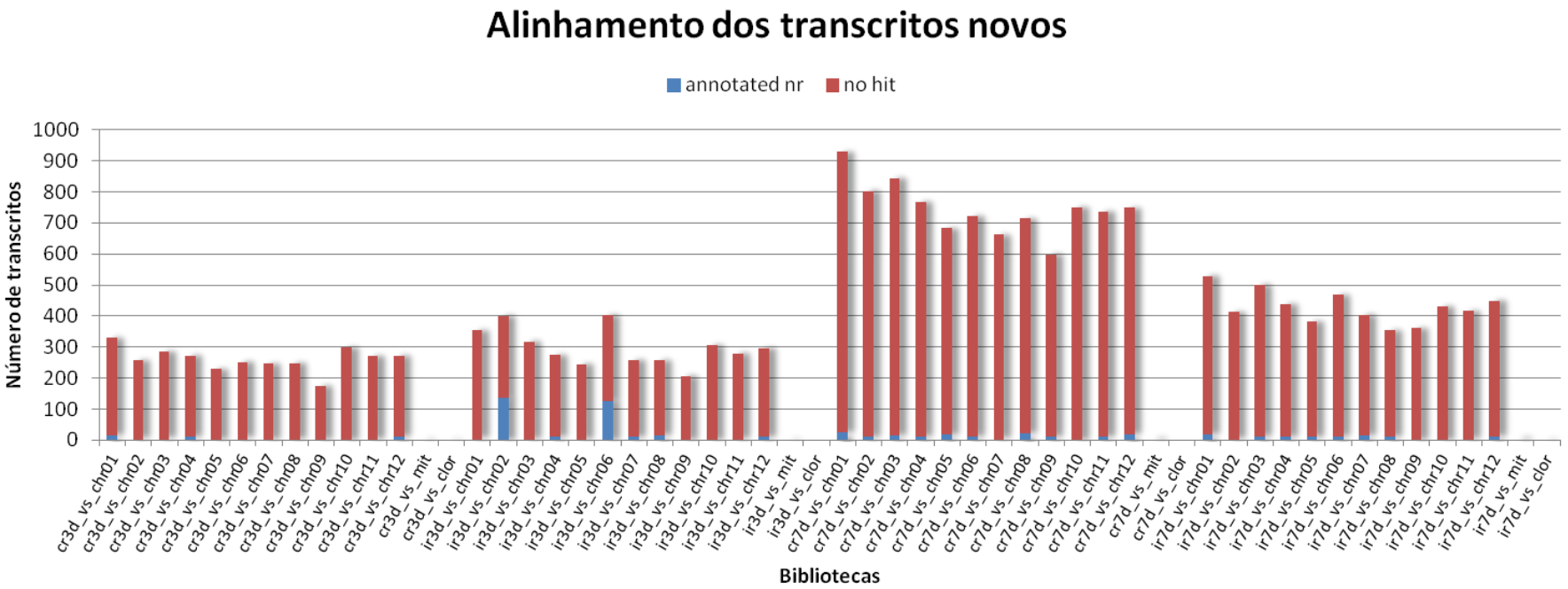


Figura 78 - Alinhamento dos transcritos novos contra o NR. Em azul, 3.31% dos transcritos obtiveram *hit*.

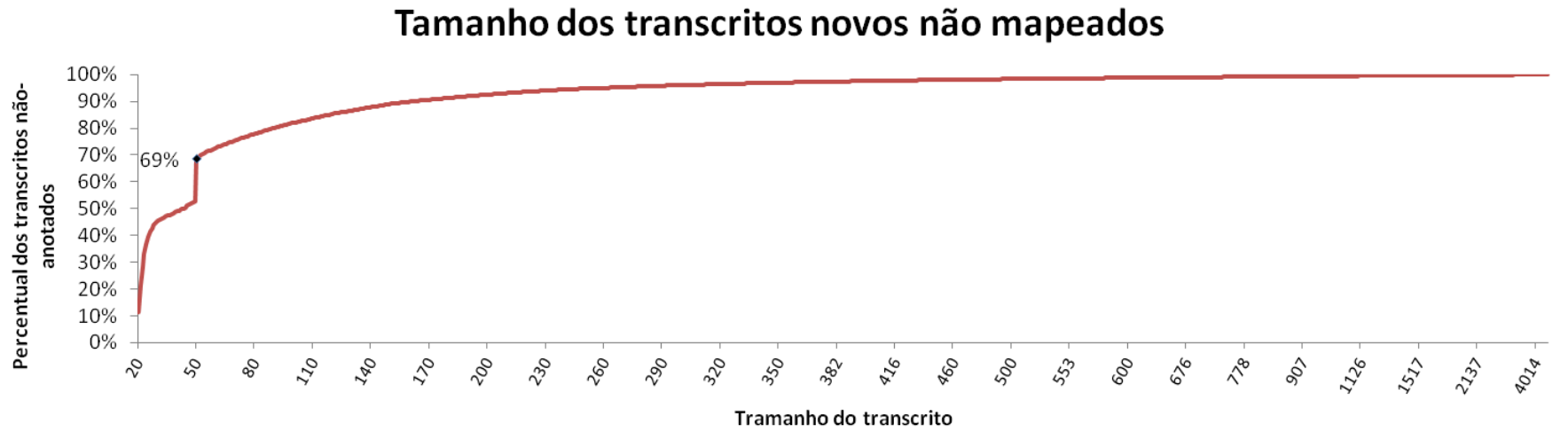


Figura 79 - Tamanho dos transcritos novos que não mapearam contra a base de dados NR.

Destes, 69% possuem até 50 pb inclusive. Portanto, 31% possuem mais de 50 pb.

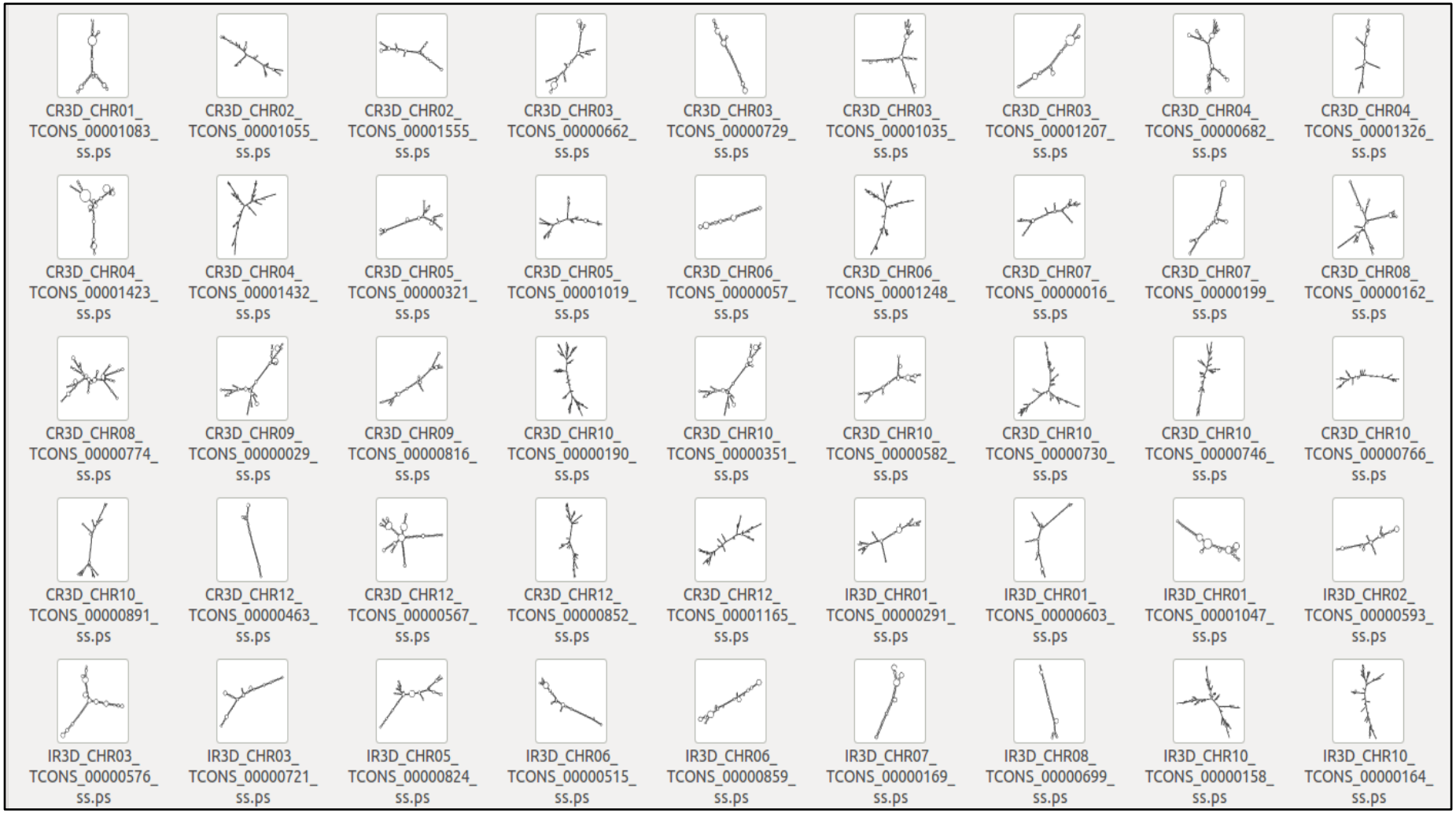


Figura 80 - Predição de estrutura secundária de transcritos não anotados, RNAFold (HOFACKER, FONTANA, *et al.*, 1994), correspondente às estruturas de precursores de microRNAs similares àquelas depositadas no banco de dados miRBase.

7. DESAFIOS

Este tópico foi escrito com intuito de citar os desafios e as soluções encontradas durante este trabalho para que possa auxiliar outros pesquisadores em análise de transcriptoma por RNA-Seq. Seguem os itens:

1. A análise utilizou 526.795.261 *reads* brutas destas, 328.409.635 foram usadas após o filtro de *trimming* contra um genoma de referência contendo 12 cromossomos mais as sequencias representativas do cloroplasto e mitocôndria. Trata-se de um desafio computacional intenso, uma vez que durante o mapeamento, muitas vezes o fornecimento de energia para o cluster era interrompido, onde o *nobreak* não mantinha o *cluster* em funcionamento por tempo suficiente até o fornecimento ser restabelecido. Neste caso, tornou-se necessário reiniciar todo o mapeamento. Outro desafio organizacional é a correta distribuição do *input* a ser analisado no sistema de arquivos do cluster, uma vez que se torna impraticável a transferência de todo o conjunto de dados repetidamente entre os computadores utilizados. Outro desafio, é com relação aos backups realizados; de fundamental importância para garantir o resultado das análise, salvando assim tempo computacional caso algum problema nos servidores venha a causar a perda dos dados.
2. Foi necessário criar um arquivo próprio referente as anotações das características genômicas das organelas (mitocôndria e cloroplasto). Como estes arquivos, durante o desenvolvimento deste trabalho, não estavam disponibilizados pelo NCBI, utilizamos a tabela de características encontrada neste site, para criar manualmente nossos arquivos no formato GFF que fossem passíveis de serem incluídos em nosso pipeline de análise.
3. Algumas incompatibilidades foram identificadas durante a etapa de alinhamento das reads versus o genoma de referência. Notamos que a anotação contida na primeira coluna de uma característica do arquivo GFF precisa ser idêntica àquela do título das sequencias representativas à esta característica no arquivo FASTA. Assim, após várias tentativas mal-sucedidas, pudemos alterar manualmente esta discordância e proceder com as análises de maneira correta.

4. As *reads* criadas com o SOLiD obedecem um critério exclusivo do protocolo de análise, que faz com que estas sejam sequenciadas para a mesma fita do DNA codificador. A questão é que, por padrão, o Bowtie/TopHat alinha as reads fornecidas contra as 2 fitas do genoma de referência. Assim, tivemos que utilizar o parâmetro –library-type, fazendo com que apenas reads direcionadas no mesmo sentido (fita específica) sejam montadas em uma unidade de transcrição, evitando assim que alinhadas em ambas fitas montem um transcrito biologicamente inexistente. Os resultados são mascarados, principalmente no caso do sequenciador SOLiD que é fita específica.
5. Os genes de arroz possuem pelo menos 3 ID's diferentes em bancos públicos, para tanto, foi necessário criar no banco de dados local uma tabela de conversão dos mesmos.
6. Nos primeiros testes, as saídas do Cufflinks não continham o nome do gene, o que não pode acontecer visto que era necessário saber qual gene se trata na análise. A solução foi utilizar o módulo do gffread, disponível na distribuição do Cufflinks para corrigir o problema com o arquivo de anotação correto. Posteriormente, foi criada uma coluna extra no arquivo de anotação (GFF), com os nomes dos genes, para leitura no gffread. Maiores informações podem ser encontradas na própria referência do Cufflinks (http://cufflinks.cbcb.umd.edu/gff.html).

8. PUBLICAÇÕES

A Dissertação aqui apresentada, como requisito parcial à obtenção do título de Mestre em Bioinformática pela Universidade Federal do Paraná (UFPR), viabilizou a publicação dos trabalhos listados abaixo:

Resumos publicados em anais de congressos:

LOPES, K. P. ; COVRE, R A ; Brusamarello-Santos ; Wassem R ; SOUZA, E M ; PEDROSA, F O ; Barbosa-Silva . **Initial mapping of rice cdna short-sequences created by rna-seq.** In: 7th Conference of the Brazilian Association for Bioinformatics and Computacional Biology, 2011, Florianópolis - SC. Transcriptomes, Proteomics, 2011.

LOPES, K. P. ; COVRE, R A ; Brusamarello-Santos ; SOUZA, E M ; Wassem R ; PEDROSA, F O ; Barbosa-Silva . **Identification of rice alternatively spliced and new transcripts in response to diazotrophic inoculation using RNA-Seq data**. In: 8th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (X- Meeting), 2012, Campinas - SP. Transcriptomes and Proteomics, 2012.

LOPES, K. P. ; COVRE, R A ; Brusamarello-Santos ; Wassem R ; SOUZA, E M ; PEDROSA, F O ; Barbosa-Silva . **RNA-Seq based identification of alternative splicing in transcription start sites of rice mRNAS.** In: 8th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (X- Meeting), 2012, Campinas - SP. Transcriptomes and Proteomics, 2012.

Menção Honrosa durante o X-meeting 2012 (Conferência Internacional da Associação Brasileira de Bioinformática e Biologia Computacional), realizado em Campinas - SP no período de 14 a 17 de Outubro, para o trabalho intitulado: "*RNA-Seq based identification of alternative splicing in transcription start sites of rice mRNAS.*" Área: Transcriptômica e Proteômica.

O artigo para submissão em revista da área está sendo escrito. Outros dados referentes aos genes diferencialmente expressos e as categorias biológicas enriquecidas também foram criados, porém estes foram compartilhados com os pesquisadores do Depto. de Bioquímica da UFPR, equipe interessada no estudo da expressão diferencial de genes em arroz quando este encontra-se inoculado com a bactéria *H. seropedicae* (Brusamarello *et al*., dado não publicado).

REFERÊNCIAS

ALTSCHUL, S. F. et al. Basic Local Alignment Search Tool. **Journal of Molecular Biology**, v. 215, p. 403-410, 1990.

BALDANI, J. I. et al. Characterization of Herbaspirillum seropedicae gen. nov., sp. nov., a Root-Associated Nitrogen-Fixing Bacterium. **International Journal of Systematic Bacteriology**, EMBRAPA-Programa Nacional de Pesquisa em Biologia do Solo, Seropédica, 23851, Rio de Janeiro, Brazil, v. 36, p. 86-93, Jan. 1986.

BRAWAND, D. et al. The evolution of gene expression levels in mammalian organs. **Nature**, v. 478, n. 343, p. 8, October 2011. ISSN DOI:10.1038/nature10532.

BRAZMA, A. et al. Minimum information about about a microarray experiment (MIAME)_towad standards for microarray data. **Nature genetics**, v. 29, December 2001.

BRUSAMARELLO, L. C. C.; WASSEM, R.; SOUZA, E. M. Sequenciamento e análise de seqüências expressas (ESTs) de arroz (oryza sativa) inoculado com herbaspirillum seropedicae, Curitiba, 2007.

CARVALHO, M. C. D. C. G.; SILVA, D. C. G. D. Sequenciamento de DNA de nova geração e suas aplicações na genômica de plantas. **Ciência Rural**, Santa Maria, v. 40, p. 10, Março 2010. ISSN ISSN 0103-8478.

CELOTTO, A. M.; GRAVELEY, B. R. Alternative Splicing of the Drosophila Dscam Pre-mRNA Is Both Temporally and Spatially Regulated. **Genetics**, p. 599–608, October 2001.

COOPER, G. M. Recombinant DNA. **NCBI**, 2000. Disponivel em: <http://www.ncbi.nlm.nih.gov/books/NBK9950/>. Acesso em: 04 December 2012.

CUMMINGS, C.; SCIENTIST, S. Microbial genome sequencing on the Applied BiosystemsSOLiD and Ion Torrent PGM - NGS platforms, 2012. Disponivel em: <http://www.molecularevolution.org/molevolfiles/presentations/craig_cummings_fortcollins_2011.pdf>. Acesso em: 03 December 2012.

EMBRAPA. Agência de Informação EMBRAPA - Cana de açúcar. **Ministério da Agricultura, Pecuária e Abastecimento**, 2012. Disponivel em: <http://www.agencia.cnptia.embrapa.br/gestor/cana-de-acucar/arvore/CONTAG01_31_711200516717.html>. Acesso em: 22 Outubro 2012.

GRAVELEY, B. R. Alternative splicing: increasing diversity in the proteomic world. **Trends in genetics**, v. 17, p. 100-107, February 2001.

HOFACKER, I. L. et al. RNAfold WebServer. **RNAfold**, 1994. Disponivel em: <http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi>. Acesso em: 07 December 2012.

INTERNATIONAL RICE GENOME PROJECT. The map-based sequence of the rice genome. **Nature**, v. 436, p. 8, 11 August 2005. ISSN doi:10.1038/nature03895.

ITOH, T. et al. Curated genome annoatation of Oryza sativa ssp. japonica and comparative genome analysis with Arabdopsis thaliana. **Genome Research**, p. 10, 8 January 2007. ISSN doi: 10.1101/gr.5509507.

KIDNER, C. A.; MARTIENSSEN, R. A. The developmental role of microRNA in plants. **Plant Biology**, p. 38-44, November 2004. ISSN DOI 10.1016/j.pbi.2004.11.008.

LANGMEAD, B. et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. **BMC - Molecular Biology**, March 2009. ISSN doi:10.1186/gb-2009-10-3-r25.

LEHNINGER, A. L.; NELSON, D. L.; COX, M. M. **Princípios de Bioquímica**. Tradução de Arnaldo Antônio Simões e Wilson Roberto Navega Lodi. 4. ed. São Paulo: Sarvier, 2006.

LI, P. et al. The developmental dynamics of the maize leaf transcriptome. **Nature genetics**, v. 42, p. 9, October 2010. ISSN DOI: 10.138/ng.703.

LOPES, K. P. et al. Identification of rice alternatively spliced and new transcripts in response to diazotrophic inoculation using RNA-Seq data. **Abstracts Book, X-meeting**, Campinas/ SP, October 2012.

MALEK, O.; KNOOP, V. Trans-splicing group II introns in plant mitochondria: The complete set of cis-arranged homologs in ferns, fern allies, and a hornwort. **RNA - A**

publication of the RNA society, p. 1599-1609, 24 September 1998. ISSN RNA (1998), 4:1599–1609.

MARDIS, E. Next-Generation DNA Sequencing Methods. **Annual Review of Genomics and Human Genetics**, p. 387-402, 2008.

MOORE, M. J. et al. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. **PNAS**, p. 1-6, January 2010. ISSN pnas.0907801107.

MORERA, D. et al. RNA-Seq Reveals an Integrated Immune Response in Nucleated Erythrocytes. **PLoS One**, p. 1 - 9, 27 October 2011.

MORTAZAVI, A. et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. **Nature Methods**, July 2008. ISSN PMID: 18516045.

NAGALAKSHMI, U. et al. The transcriptional landscape of the Yeast Genome defined by RNA sequencing. **Science**, v. 320, p. 7, 2008. ISSN DOI: 10.1126/science.1158441.

NOBUTA, K. et al. An expression atlas of rice mRNAs and small RNAs. **Nature biotechnology**, p. 5, March 2007. ISSN doi:10.1038/nbt1291.

PEDROSA, F. O. et al. Genome of Herbaspirillum seropedicae Strain SmR1, a Specialized Diazotrofic Endophyte of Tropical Grasses. **Plos Genetics**, p. 10, 12 May 2011. ISSN doi:10.371/journal.pgen.1002064.

PROSDOCIMI, F. Montagem de Genoma e Transcriptoma, Rio de Janeiro, 2012. Slides. Acompanha texto.

PUTNEY, S. D.; HERLIHY, W. C.; SCHIMMEL, P. A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. **Nature**, p. 718-721, 21 April 1983. ISSN doi:10.1038/302718a0.

RICE ANNOTATION PROJECT CONSORTIUM. The Rice Annotation Project Database (RAP-DB): 2008 update. **Nucleic Acids Research**, v. 36, p. 6, 17 December 2008. ISSN doi:10.1093/nar/gkm978.

ROBINSON, J. T. et al. Integrative Genomics Viewer. **Nature Biotechnology**, v. 29, p. 24-26, January 2011.

ROTHBERG, J. M. et al. An integrated semiconductor device enabling non-optical genome sequencing. **Nature**, v. 475, p. 5, 21 July 2011.

SANGER, F.; NICKLEN, S.; COULSON, A. R. DNA sequencing with chain-terminating inhibitors. **PNAS**, v. 74, October 1977.

SUTCLIFFE, J. G. et al. Common 82-nucleotide sequence unique to brain RNA. **Proc Natl Acad Sci U S A**, v. 79, p. 4942-4946, August 1982.

TRAPNELL, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. **Nature Protocols**, v. 7, p. 562-578, March 2012. ISSN doi:10.1038/nprot.2012.016.

TRAPNELL, C.; PACHTER, L.; SALZBERG, L. S. TopHat: discovering splice junctions with RNA-Seq. **Oxford Journals, Bioinformatics**, v. 25, p. 1105-1111, March 2009. ISSN doi:10.1093/bioinformatics/btp120.

VENTER, J. C. et al. The Sequence of the Human Genome. **Science**, p. 1304-1351, 2001.

WANG, Z.; GERSTEIN, M.; SNYDER, M. RNA-Seq: a revolutionary tool for transcriptomics. **Nature Review Genetics**, p. 16, January 2009. ISSN doi: 57-63.doi:10.1038/nrg2484.

WATSON, J. D. et al. **Biologia Molecular do Gene**. Tradução de Luciane Passaglia. 5ª Edição. ed. Porto Alegre: Artmed, 2006. ISBN ISBN 85-363-0684-X.

YAMASHITA, R. et al. Genome-wide Characterization of Transcriptional Start Sites in Humans by Integrative Transcriptome Analysis. **Genome Research**, March 2011. ISSN doi:10.1101/gr.110254.110.